charles river analytics

Charles River Analytics Inc.
625 Mount Auburn St.
Cambridge, MA 02138
www.cra.com

p: 617.491.3474
f: 617.868.0780

20 August 2015

Dr. Harold Hawkins
ONR Code 341
Office of Naval Research
875 North Randolph SL
Arlington. VA 22203-1995

Reference: US Navy Contract N00014-12-C-0653: "The Model Analyst's Toolkit: Scientific Model Development, Analysis, and Validation"
Charles River Analytics Contract No. C12186

Subject: Contractor's Final Report
Reporting Period: 20-August-2012 to 20-August-2015

Dear Dr. Hawkins,

Please find enclosed 1 copy of the Final Report for the referenced contract. Please feel free to contact me with any questions regarding this report or the status of the "The Model Analyst's Toolkit: Scientific Model Development, Analysis, and Validation" effort.

Sincerely,

W. Scott Neal Reilly

W. Scott Neal Reilly
Principal Investigator

cc: Cheryl Gonzales, DCMA
Annetta Burger, ONR
Whitney McCoy, Charles River Analytics

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302 Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number

1. REPORT DATE **20 AUG 2015**	2. REPORT TYPE	3. DATES COVERED **00-00-2015 to 00-00-2015**

4. TITLE AND SUBTITLE **The Model Analyst’s Toolkit: Scientific Model Development, Analysis, and Validation**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Charles River Analytics,,625 Mount Auburn Street,,Cambridge,,MA, 02138**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a REPORT **unclassified**	b ABSTRACT **unclassified**	c THIS PAGE **unclassified**	**Same as Report (SAR)**	**56**	

Charles River Analytics

Monthly Technical Progress Report No. R12186-20
Reporting Period: August 20, 2012 to August 20, 2015
Government Contract No. N00014-12-C-0653
Charles River Analytics Contract No. C12186

The Model Analyst's Toolkit: Scientific Model Development, Analysis, and Validation

Final Report

Principal Investigator: Scott Neal Reilly

Charles River Analytics
625 Mount Auburn Street
Cambridge, MA 02138
617-491-3474

August 20, 2015

The views opinions and findings contained in this report are those of the authors and should not be construed as an official Agency position policy or decision unless so designated by other official documentation.

1. Executive Summary

The effort builds on and extends the work of the previous ONR-funded "Validation Coverage Toolkit for HSCB Models" project. The overall objectives of the research program are:

- Help scientists create, analyze, refine, and validate rich scientific models
- Help computational scientists verify the correctness of their implementations of those models
- Help users of scientific models, including decision makers within the US Navy, to use those models correctly and with confidence
- Use a combination of human-driven data visualization and analysis, automated data analysis, and machine learning to leverage human expertise in model building with automated analyses of complex models against large datasets

Specific objectives for the current effort include:

- **Fluid temporal correlation analysis.** Our objective is to design a new method for performing temporally fluid correlation analysis for temporal sets of data and implement the method as a new prototype component within the Model Analyst's Toolkit (MAT) software application.
- **Automated suggestions for model construction and refinement.** Our objective is to design and implement a prototype mechanism that learns from data how factors interact in non-trivial ways in scientific models.
- **Data validation and repair.** Our objective is to design and implement a prototype capability to identify likely errors in data based on anomalies relative to historic data and to use models of historic data to offer suggested repairs.
- **System prototyping.** Our objective is to incorporate all improvements into the MAT software application and make the resulting application available to the government and academic research community for use in scientific modeling projects.
- **Evaluation of applicability to multiple scientific domains.** Our objective is to ensure (and demonstrate) that MAT can be applied to a wide range of scientific domains by identifying and building at least one neurological and/or physiological model and analyze the associated data with MAT, making any extensions to the MAT tool that are necessary to support the analysis of such a model.

2. Overview of Problem and Technical Approach

2.1. Summary of the Problem

One of the most powerful things scientists can do is to create models that describe the world around us. Models help scientists organize their theories and suggest additional experiments to run. Validated models also help others in more practical applications. For instance, in the hands of military decision makers, human social cultural behavior

(HSCB) models can help predict instability and the socio-political effects of missions, whereas models of the human brain and mind can help educators and trainers create curricula that more effectively improve the knowledge, skills, and abilities of their pupils.

While there are various software tools that are used by the scientific community to help them develop and analyze their models (e.g., Excel, R, Simulink, Matlab), they are largely so general in purpose (e.g., Excel, R) or so focused on computational models in particular (e.g., Simulink, Matlab), that they are not ideal for rapid model exploration or for use by non-computational scientists. They also largely ignore the problem of validating the models, especially when the models are positing causal claims as most interesting scientific models do. To address this gap, Charles River Analytics undertook the "Validation Coverage Toolkit for HSCB Models" project with ONR. Under this effort, we successfully designed, implemented, informally evaluated, and deployed a tool called the Model Analyst's Toolkit (MAT), which focused on supporting social scientists to visualize and explore data, develop causal models, and validate those models against available data (Neal Reilly, 2010; Neal Reilly, Pfeffer, Barnett et al., 2011, 2010).

As part of the development of the MAT tool, we identified four important extensions to that research program that would further support the scientific modeling process:

- Correlation analyses are still the standard way of identifying relationships between factors in a model, but correlations are fundamentally flawed as a tool for analyzing potentially causal or predictive relationships as they assume instantaneous effects. Even performing correlation analyses with a temporal offsets between streams of data is insufficient as the temporal gap between the causal or predictive event and the following event may not be the same every time (either because of variability in the system being modeled or because of variability introduced by a fixed sampling rate). What we need is a novel way of evaluating the true predictive power across streams of data that can deal with fluid offsets between changes in one stream of data and follow events in the other stream of data.
- Modeling complex phenomena is a fundamentally difficult task. Human intuition and analysis is by far the most effective way of performing this task, but even humans can be overwhelmed by the complexity of modeling the systems they are studying (e.g., socio-political system, human neurophysiology). Automated tools, while not especially good at generating reasonable scientific hypotheses, *are* extremely good at processing large amounts of data. We believe there is an opportunity for computational systems to enhance human scientific inquiry. Under the "Validation Coverage Toolkit for HSCB Models" project, we demonstrated how automated tools could help human scientists to analyze and validate their models against data. We believe a similar approach can be used to help suggest modifications to the human-built models to make them better match the available data. To be useful, however, such automated analyses will need to be rich enough to suggest subtle data interactions that are most likely to be missed by the human scientist. For instance, correlations (especially correlations that take into account fluid temporal displacements) could be used to identify likely

relationships between streams of data, but such an approach would miss complex, non-linear relationships between interrelated factors that cannot be effectively analyzed with simple two-way correlations. For instance, if crime waves are associated with increases in unemployment *or* drops in police presence, that would be hard to identify with a correlation analysis. We need richer automated data analysis techniques that can extract complex, non-linear, multi-variable relationships between data if we are to effectively suggest model improvements to human scientists.

- Even if a scientific model is sound, if the data sets provided as inputs to the model are unreliable, the results of the model are still suspect. And, unfortunately, data will often be wrong. For instance, HSCB surveys are notoriously unreliable and biased for a variety of reasons, and neurological and physiological data can be corrupted by broken or improperly used sensors. If it were possible to identify when data was unreliable and, ideally, even repair the data, then the models that are using the data could once again be effectively used.
- The MAT tool we developed under the “Validation Coverage Toolkit for HSCB Models” project was focused primarily on assisting social scientists in the analysis, refinement, and validation of HSCB models. In parallel with that effort, however, we also took an opportunity to apply MAT to evaluating neurological and physiological data under the DARPA-funded CRANIUM (Cognitive Readiness Agents for Neural Imaging and Understanding Models) program. We discovered the generality of the MAT tool makes it potentially applicable to a great number of different scientific domains. MAT proved to be a useful, but peripheral tool, in CRANIUM. We believe MAT could be applied to a broader suite of scientific modeling problems than it has been so far.

2.2. Summary of our Approach

To address these identified gaps and opportunities, we extended MAT’s support for model development, analysis, refinement, and validation; enhancing MAT to analyze and repair data; and demonstrating MATs usefulness in additional scientific modeling domains. Our approach encompasses the following four areas, which correspond to the four gaps/opportunities identified in the previous section:

- **Temporally Fluid Correlation Analysis.** We are designed new methods to perform Temporally Fluid Correlational Analysis on temporal sets of data, and we implemented the methods as a new component within the MAT software application. The version of MAT at the beginning of the new effort supported correlation analysis for temporally offset data; it shifts the two data streams being compared by a fixed offset that is based on the sampling rate of the data (i.e., data that is sampled annually will be shifted by one year at a time), performs a standard correlation on the shifted data, plots the correlation value against the amount of the offset, and then repeats the process for the next offset amount. If two data streams are shifted by a fixed offset (e.g., changes in one stream are always followed by a comparable value in the other stream after a fixed time), then this method will find that offset. Under the current effort, we expanded on this capability to support fluid temporal shifts within the data streams. That is, we make it possible to identify when the temporal offset between the change in the first data stream and its effect in the second stream is not a static amount of time.

- **Automated suggestions for model construction and refinement.** We designed and implemented a mechanism to learn how factors interact in non-trivial ways in scientific models. In particular, we developed a method for learning disjuncts, conjuncts, and negations. This mechanism starts with the model developed by the scientist user and make recommendations for possible adjustments to make it more complete by performing statistical data mining and machine learning.
- **Data validation and repair.** Recognizing that data contains errors is plausible once we understand the relationships between data sets. That is, if we are able to develop models of the correlations between sets of data, then we can build systems that notice when these correlations do not hold in new data, indicating possible errors in data. For instance, if we know that public sentiment tends to vary similarly between nearby towns, then when one town shows anomalous behavior, we can reasonably suspect problems with the data. There might be local issues that cause the anomaly, but it is, at least, worth noting and bringing to the attention of the user of the data and model. As MAT is designed to help analyze models and recognize inter-data relationships, it is primed to perform exactly this analysis. Existing methods perform similar types of analysis for environmental data (Dereszynski & Dietterich, 2007, 2011). For instance, a broken thermometer can be identified and the data from it even estimated by looking at the temperature readings of nearby thermometers, which will generally be highly correlated.
- **Application to multiple scientific modeling domains.** To ensure (and demonstrate) that MAT can be applied to a wide range of scientific domains, we identified and built a number of models from differing scientific domains and analyzed the associated data with MAT, making any extensions to the MAT tool necessary to support the analysis of such a model. The initial MAT effort focused on HSCB models; by focusing this effort on harder-science models at much shorter time durations, we believe we have demonstrated an interesting range of applications of the MAT tool.

3. Accomplishments of the Program

In this Section, we review the accomplishments of the effort. In Section 3.1, we describe our results in creating temporally fluid causal analysis techniques; in Section 3.2, we present our results on automated model recommendations; in Section 3.3, we describe our results in data validation; in Section 3.4, we describe out efforts to apply MAT to a variety of scientific domains. Finally, we made a number of other improvements to the usability and efficiency of the software that are not directly tied to other objectives; these are summarized in Section 3.5. Our transition and marketing efforts are described in Section 3.6.

3.1. Temporally Fluid and Other Causal Model Analysis Techniques

Our goal for this task is to identify algorithms and tools that can help scientists to analyze causal relationships in their data. Under the previous MAT program, we developed the ability to do correlation analyses of temporally offset data. The idea was that, if causes precede effects, then we might find cases where the purported cause with the data shifted forward in time has a higher correlation than if there was no shift in the two data series. That is, offset correlations may help identify when there is a temporal lag between two data series, which is evidence of a potentially causal relationship between the two series. The initial implementation, however, only worked for static offsets and we expect that the time between the cause and effect might vary slightly between instances of causes and effects.

Under the current effort, we are developing methods so that this offset can be more flexible, since offsets in many scientific domains are rarely fixed. To address these cases, we are exploring more advanced methods for validating causal relationships in models. This validation can help researchers produce more robust models of complex systems to facilitate the testing of dependencies that might otherwise be missed, assumed away, or taken for granted. We reviewed the following four methods:

- Granger Causality (GC) – This statistical method comes from work in econometrics and was designed to attempt to find predictive patterns in temporal data.
- Dynamic Time Warping (DTW) – This is inspired by work in gait recognition, where the same gait must be recognized even when the subject is slowing down or speeding up.
- Convergent Cross-Mapping (CCM) – This is a relatively new method that is effective at identifying cyclical causal patterns, such as found in predator-prey systems.
- Feature-Based Pattern Detection – This approach first pulls out "interesting" features from the data (e.g., large drops, spikes) and looks for patterns between those features across multiple datasets. This approach can identify causal and predictive patterns even where there is little to no statistical correlation in the data values.

The following sections discuss the first three of these methods. The fourth is described in more detail in Section 3.2.3.

3.1.1. Granger Causality for Validating Dependencies

Granger Causality (GC) was originally introduced for economic models (Granger, 1980, 1969) to help deal with the problem of temporal offsets. It can, however, be adapted as a validation test for causality in socio-cultural data. Granger causality makes two assumptions: (1) the effect does not precede the cause, and (2) the causal variable provides information about the effect that is otherwise unavailable.

Definition 1 (Granger Cause). *The temporal variable X* ***Granger causes*** *temporal variable Y iff* $P(Y_t \mid Y_{t-1}^{t-L}) \neq P(Y_t \mid Y_{t-1}^{t-L}, X_{t-1}^{t-L})$ *where L is the maximum time lag,* a_i*,* b_j *are parameters in a linear combination,* ϵ_1*,* ϵ_2 *are error terms, and*

$$P(Y_t \mid Y_{t-1}^{t-L}) \; = \; \sum_{l=1}^{L} a_l \, Y_{t-l} + \epsilon_1 \tag{1}$$

$$P(Y_t \mid Y_{t-1}^{t-L}, X_{t-1}^{t-L}) \; = \; \sum_{l=1}^{L} a_l \, Y_{t-l} + \sum_{l=1}^{L} b_l \, X_{t-l} + \epsilon_2 \tag{2}$$

Variable X is a Granger cause of Y if Y is better predicted using the histories of X and Y than just of Y alone. We can validate this relationship through hypothesis testing. If equation (2) is statistically more accurate than equation (1) using an F statistic, then a causal relationship between X and Y is valid. Figure 3-1 shows this test.

Granger Causality *F*-test Validation Procedure
H_0: X does not Granger cause Y.

1. Choose the desired significance level α and identify the critical value c.
2. Use equations (1) and (2) to compute the F-statistic where (1) is a restricted model of (2) such that all of the coefficients $b_{t-1} = b_{t-2} = ... = b_{t-L} = 0$.
3. If $F > c$, then the null hypothesis can be rejected and a causal relationship is validated.

Figure 3-1: Granger causality *F*-test validation

When implementing this metric, we found that it was sensitive to multiple linear regression, so we experimented with several regression techniques. We are currently using the algorithm from the Apache commons, and are considering allowing the user to select from multiple regression methods in future versions of MAT.

We implemented the GC metric in a Matlab prototype, then ported the Matlab prototype into our Java-based MAT application.

3.1.2. Dynamic Time Warping for Uneven Temporal Relationships

Many causal relationships are imperfectly represented in the observed data. This is particularly salient in complex socio-cultural systems where variability in human behavior produces uneven temporal delays between cause and effect. For example, lower employment rates may cause an increase in future (e.g., 6 to 12 months) crime rate. These relationships cannot be captured by standard statistical analyses (including GC), which assume a stationary process with a consistent time lag.

To validate these relationships, we borrowed and extended the dynamic time warping (DTW) algorithm from gait recognition (Salvador & Chan, 2007; Myers & Rabiner, 1981), where DTW identifies a gait from two motion curves even when a person speeds up or slows down. The DTW algorithm compares the two time series to find the optimal alignment by “warping” one series by stretching or shrinking it along its time axis.

Definition 2 (Warp Path). *Given two time series X and Y of size n and m a* ***warp path*** *W is a sequence W=w_1, w_2, …, w_K where K is the length of the path and each element w_k = (i, j) represents a mapping between point i in X with point j in Y. The optimal warp path minimizes the sum of distances between the mapped points*

$$argmin\ Dist(W) = \sum_{k=1}^{k=K} Dist(w_{ki}, w_{kj})$$

where $Dist(W)$ is the distance of warp path W and $Dist(w_{ki}, w_{kj})$ is the distance between point i in series X and point j in series Y.

A warp path identifies a series of cells in an $n \times m$ two-dimensional cost-matrix D, where each $D(i,j)$ is the minimum distance warp path that can be constructed from X up to x_i and Y up to y_i. The final entry contains the optimal path over the full series.

Because causality only impacts the future, we expanded DTW to handle the one-directional case in a new algorithm, ForwardDTW. Rather than matching points in both directions, ForwardDTW only matches the points in X with future values of Y. That is, the entry $D(i,j)$ is only computed for $i < j$. ForwardDTW enables us to use DTW to validate causal relationships—the smaller the warp distance between X and Y, the stronger the causal link. A user can specify this causality threshold to determine when a relationship is considered validated, as shown in Figure 3-2.

Dynamic Time Warping Validation Procedure

1. Set a warping threshold *t*.
2. Use the ForwardDTW algorithm to compute $min\ Dist(W)$for X and Y.
3. If $min\ Dist(W) < t$, a causal relationship is validated.

Figure 3-2: Dynamic time warping validation procedure

Some advantages of DTW over other time series analyses are that it can account for missing data and compare series with different time scales or sampling frequencies. DTW is also very visual, making the results easy to interpret by human analysts.

One concern with DTW is that it is flexible and powerful enough to make even dissimilar datasets appear similar (to a point). To address this, MAT provides two methods to control and evaluate the results of the DTW algorithm. In particular, MAT provides the user the ability to limit how much warping is permissible and it displays the warped data, drawing visual attention to the shift, compression, and expansion regions so that the user can easily see what the warping is doing to match the curves.

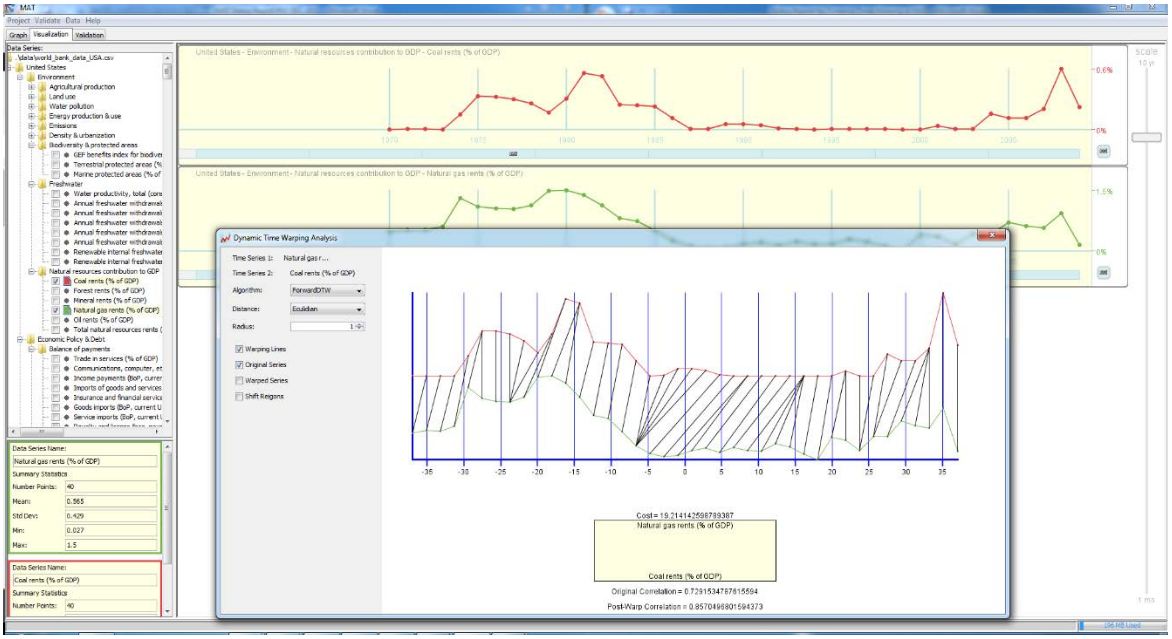
DTW uses a cost matrix to determine the optimal "warping" between two data sets. Warping is defined as the mapping between data points in one series to data points in another that minimizes the sum of distances between the mapped points. One major advantage of DTW compared to other algorithms used for time series comparison is that it can account for missing data points and can compare two data sets that use different time scales or sampling frequencies by compressing or expanding certain areas of the data set.

One limitation of the DTW algorithm is the efficiency in both time and space used by the algorithm. Because it compares all the points to each another in a cost matrix, the algorithm runs in quadratic time and uses quadratic space in relation to the size of the input data set. When comparing large data sets, this cost can make the algorithm unusable. However, academic research into DTW optimization has resulted in a breakthrough to reduce this cost, providing a linear time approximation algorithm (FastDTW) that performs far better than previous approximations (Chan & Salvador, 2007).

To quickly incorporate the DTW algorithm into MAT, we used the free, open source Java library included with the paper describing the algorithm (Chan et al., 2007). Both the regular DTW and FastDTW algorithms are implemented in this open-source library and have been incorporated into MAT.

Because we envision that the DTW algorithm in MAT will be used not only to compare two time series, but also to analyze causal relationships, we slightly modified the original DTW algorithm. To match time series together, the original DTW algorithm matches points in either direction, allowing some points in the series to shift forward as well as backward to better match up. However, a backward shift in time is nonsensical in this use case and would mislead the user, so we removed this ability. We named this algorithm ForwardDTW.

The functionality of DTW implemented under the previous MAT effort enables users to select two time series for comparison, displaying the traditional warping lines between the two series, as shown in Figure 3-3. In this case, we are comparing natural gas rents (green) with coal rents (red).


MAT
Project Validate Data Help
Graph
Visualization
Validation
Data Series:
.\data\world_bank_data_USA.csv
United States
Environment
Agricultural production
Land use
Water pollution
Energy production & use
Emissions
Density & urbanization
Biodiversity & protected areas
GEF benefits index for biodiver
Terrestrial protected areas (%
Marine protected areas (% of
Freshwater
Water productivity, total (cons
Annual freshwater withdrawals
Annual freshwater withdrawals
Annual freshwater withdrawals
Annual freshwater withdrawals
Annual freshwater withdrawals
Renewable internal freshwater
Renewable internal freshwater
Natural resources contribution to GDP
☑ Coal rents (% of GDP)
□ Forest rents (% of GDP)
□ Mineral rents (% of GDP)
☑ Natural gas rents (% of GDP)
□ Oil rents (% of GDP)
□ Total natural resources rents (
Economic Policy & Debt
Balance of payments
Trade in services (% of GDP)
Communications, computer, et
Income payments (BoP, currer
Imports of goods and services
Insurance and financial service
Goods imports (BoP, current U
Service imports (BoP, current U
United States - Environment - Natural resources contribution to GDP - Coal rents (% of GDP)
0.6%
0%
1970
1975
1980
1985
1990
1995
2000
2005
United States - Environment - Natural resources contribution to GDP - Natural gas rents (% of GDP)
1.5%
0%
Scale
10 yr
1 mo
Dynamic Time Warping Analysis
-35
-30
-25
-20
-15
-10
-5
0
5
10
15
20
25
30
35
Natural gas rents (% of GDP)
Coal rents (% of GDP)

Data Series Name: Natural gas rents (% of GDP)
Summary Statistics
Number Points: 40
Mean: 0.565
Std Dev: 0.429
Min: 0.027
Max: 1.5

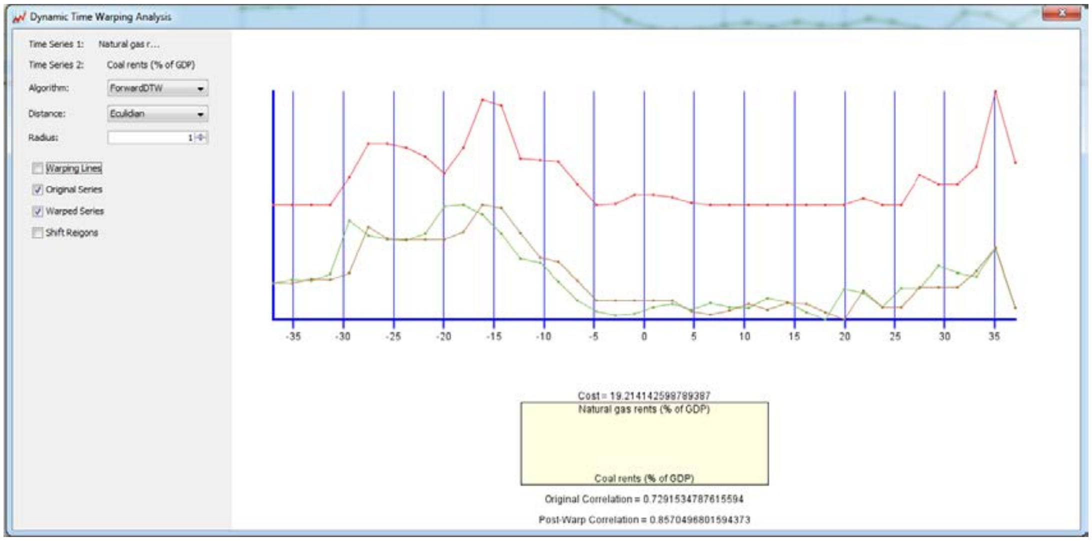
Data Series Name: Coal rents (% of GDP)
Summary Statistics
Number Points: 40

Time Series 1: Natural gas r...
Time Series 2: Coal rents (% of GDP)
Algorithm: ForwardDTW
Distance: Eculidian
Radius: 1
☑ Warping Lines
☑ Original Series
□ Warped Series
□ Shift Reigons

Cost = 19.214142598789387
Original Correlation = 0.7291534787615594
Post-Warp Correlation = 0.8570496801594373

Figure 3-3: MAT timewarping analysis capability

The user can also include the resulting time series if the warping process shifts, compresses, and expands the time series to better match the other, as shown in Figure 3-4. In this case, we shift the green line to create the brown line, which better matches the red line. The user can hide the lines to better view the shift and is given the pre-warp and post-warp correlation values for the two data sets.



Dynamic Time Warping Analysis

Time Series 1: Natural gas r...

Time Series 2: Coal rents (% of GDP)

Algorithm: ForwardDTW

Distance: Eculidian

Radius: 1

□ Warping Lines

☑ Original Series

☑ Warped Series

□ Shift Reigons

Cost = 19.214142598789387

Natural gas rents (% of GDP)

Coal rents (% of GDP)

Original Correlation = 0.7291534787615594

Post-Warp Correlation = 0.8570496801594373

Figure 3-4: Timewarped data in MAT

One concern with DTW is that it can make dissimilar data sets appear similar. By providing users with some metric or threshold based on the cost of the warp and improvement in correlation, we may be able to prevent such false positives in the majority of cases.

3.1.3. Convergent Cross-Mapping for Dynamic Feedback Models

Many social systems contain feedback relationships, where dependency between variables is bi-directional—declining economic output may increase levels of political violence, which further depresses the economy, etc. These relationships are extremely difficult to validate using standard approaches. To analyze cyclic causality, we used the CCM algorithm (Sugihara, May, Ye et al., 2012). CCM was first introduced in biology to model predator/prey systems, but can be adapted to model the interrelationships in other types of scientific data.

To use CCM, we derived a set of vectors for variables X and Y called *shadow attractor manifolds* to represent a topological projection of the underlying dynamic system.

Definition 3 (Shadow Attractor Manifold). *For a time series variable X, the* ***shadow attractor manifold*** *M_X consists of points* $x(t) = (X(t), \big(X(t-\tau), X(t-2t), \ldots, X(t-E\tau)\big)$, *where τ is a sampling time lag and E is the manifold dimension.*

For subsets of the time series X and Y of length L, we can construct manifolds M_X and M_Y. CCM then determines how well local "neighborhoods"—small regions of M_X —correspond to neighborhoods in M_Y. If X and Y are causally linked, there is a one-to-one mapping between points in M_X and M_Y . To compute this cross mapping, we use a neighborhood in M_X to predict the values of contemporaneous points in M_Y and compute the correlation ρ between the predicted values $\boldsymbol{\hat{Y}(t)}$ and the real values $Y(t)$. If a causal relationship exists, predictions of the state of Y from X (and vice versa) improve asymptotically as the amount of data (L) increases; that is, the mapping of X and Y will converge to perfect predictability $\rho = 1$. Figure 3-5 lays out this process.

Convergent Cross Mapping Validation Procedure

1. Randomly choose segments of length L from X and Y.
2. Construct shadow attractor manifolds M_X and M_Y for X_L and Y_L.
3. Compute the cross mapping ρ between M_X and M_Y in both directions.
4. If ρ converges toward 1as L increases, then there is a causal link.

Figure 3-5: Convergent cross mapping validation procedure

In Figure 3-6, the red time series (center) is generated from the blue reference series (left); the green series (right) is generated from an independent set of parameters.

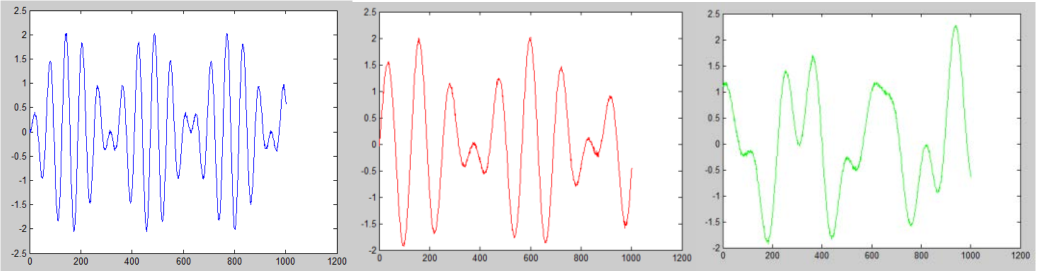


Figure 3-6: Three synthetic time series used to validate the CCM implementation

In Figure 3-7, the plot on the left illustrates CCM output for the unrelated blue and green time series; the plot on the right illustrates asymptotic prediction accuracy as a function of observation length L on the blue and red time series.

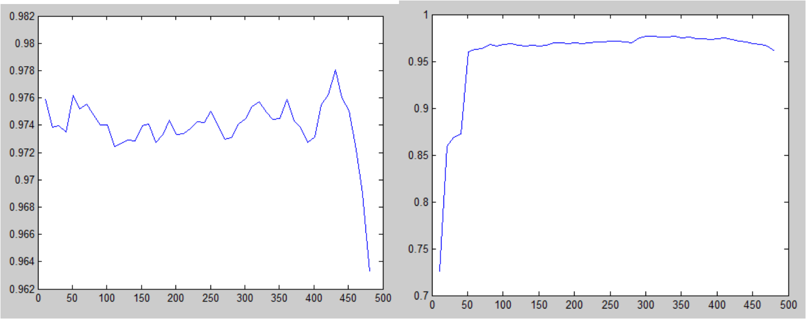


Figure 3-7: CMM output for the synthetic time series

3.1.4. New Causal Analysis Method: Beta Process Predictor Analytical Method

We have developed an additional causality analysis methodology provisionally called a *beta process predictor*. The idea behind this method is to model causal distributions as beta distributions. Most causal or statistical methods make a Gaussian assumption about the distribution of data, which is reasonable if each event is fully independent. In practice, however, real data tends to follow beta distributions because individual data events are not independent of each other.

The method works by examining for each effect-hypothesis data point every possible percent change in the data at different offsets leading up to that point. The best match is added as the preferred offset for that point. For example, if the effect shows a percent change of +15% and there are three causal points at offsets of 1 month previously, 2 months, and 3 months with values of +5%, +17% and +25%, then an offset of 2 months will be chosen (because the difference between 15 and 17 is the least among the three candidates). This results in an *offset distribution*. For example, if we repeat this process for 1000 points in the effect dataset, then we might have 43 points at -5 months, 117 points at -4 months, 175 points at -3 months, 511 points at -2 months, and 223 points at -1 months and smaller numbers beyond -6 months back. This offset distribution is fitted to a beta distribution which yields an alpha and beta parameter which fully characterize a beta distribution. We can then measure the peakedness, or kurtosis of the resulting distribution by the equation:

$$\frac{6\left[\alpha^3+\alpha^2(1-2\beta)+\beta^2(1+\beta)-2\alpha\beta(2+\beta)\right]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)}$$

The sharper this peak, the more likely the causal relationship is taken to be and vice versa. If there is no causal relationship, we would expect the distribution to be flat (kurtosis = 0). As a final step we perform this analysis both ways, for A→B and B→A. The ratio of the higher value over the lower is taken to be the degree of causality.

We have tested this metric against synthetic data by creating fully causal series and injecting differing amounts of random noise into data, then testing to see whether the metric can detect the causality. Figure 3-8 shows our preliminary results and that Beta does outperform Granger analysis for differentiating the direction of causal links for moderate amounts of noise.

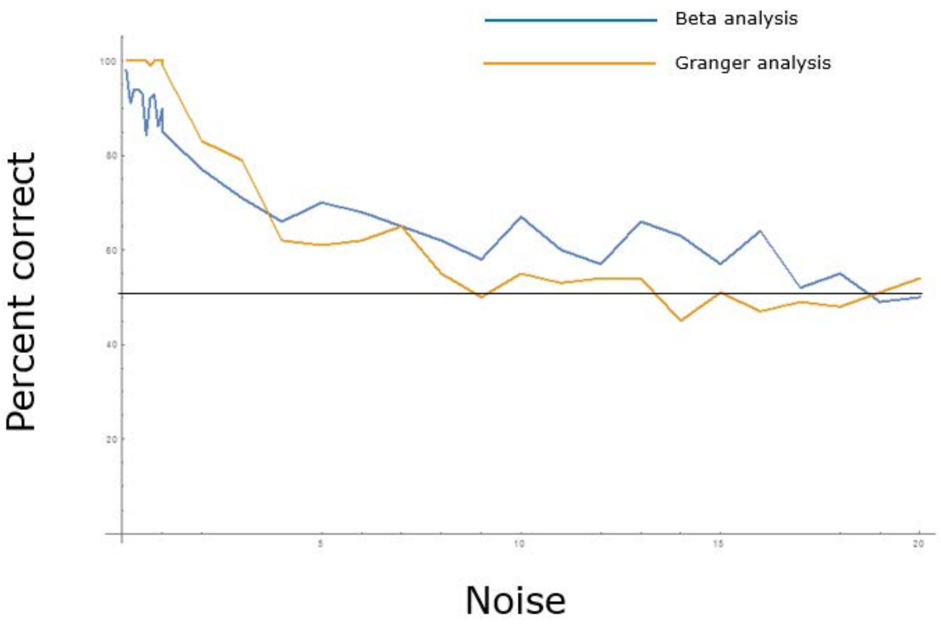


Figure 3-8. Comparison of Beta and Granger causal analysis performance

3.1.5. Integrated Causal Analysis Report

We are developing an integrated causality analysis reporting feature. Previously users of MAT would have to use different individual analytical techniques to formulate their own specialized view of indicators which could suggest a causal relationship between one or more data series. The new functionality combines multiple causality-related analytical tool calculations into a single report.

To access the new report, the user selects two series in the MAT Data View and right clicks to access the normal context menu, which now has a new choice "Causality Analysis Report." Activating the report runs the analysis and pops up the report in a dialog box. The analysis automates various parameterizations which the user would normally have to perform manually. For example, the analysis examines different temporal offsets (e.g., for Granger Causality analysis) and determines which is most likely, then uses causality metrics at those particular offsets to determine the degree of causality.

3.1.6. Evaluation of Causality Analysis Techniques

Synthetic Data-Based Evaluation

We implemented and evaluated complementary approaches to detecting causal relationships between two time series. In this section, we compare the results obtained by Granger Causality (GC) and Dynamic Time Warping (DTW) approaches for detecting causality, and discuss the importance of selecting a proper representation for the underlying data.

A qualitative comparison of GC and DTW methods on World Bank data indicates that both methods can identify potential causal (or correlative or predictive) relationships between time series. Table 1 illustrates the top 20 World Bank time series likely to share a causal relationship with the Natural Gas Rents records, as determined by DTW cost scores.

Table 2 illustrates the same results obtained using GC. Results are shown for comparisons between raw data, scaled data, and scaled and de-trended data series.

As GC inherently computes a normalized correlation score, there is no difference between raw and scaled datasets, and we omit GC results on raw data. Results are color-coded based on a series' estimated relevance to the query series—series likely to be correlated are left white; series that may be indirectly correlated are highlighted gray; series deemed unlikely to be causally correlated are highlighted yellow. A graph capturing the causal likelihood score across the top 100 series is included in each column (a lower score indicates a stronger likelihood of correlation/causation). As we expect few time series to truly exhibit causal relationships, we expect this curve to exhibit an asymptotic slope, in which a small number of time series are closely related to the query data, and similarity to other data rapidly decreases.

We found that DTW is particularly sensitive to significant differences in scale between time series, due to the underlying distance-based matching metric. This can be observed both from the quality of resulting matches, as well as from the shallow shape of the score curve. As a result, datasets must be appropriately normalized before performing analysis using DTW. To enable direct comparison between time series using DTW, we normalize datasets to the unit standard deviation.

We also found that certain time series may inherently follow long-term trends that may be present across many time series (e.g., economies tend to grow as a general trend). These trends may conceal relationships between time series that may have significant short-term impacts. To account for this possibility, we applied a linear de-trending process to each time series. This process noticeably impacts the rankings obtained using both DTW and GC techniques; in the case of DTW, this process appears to improve the overall quality of the returned results. However, as de-trending may not be appropriate for all datasets, we plan to implement this function as an optional step in the feature extraction process.

Table 1: Automatic ranking of likely causal relationships between time series, using dynamic time warping

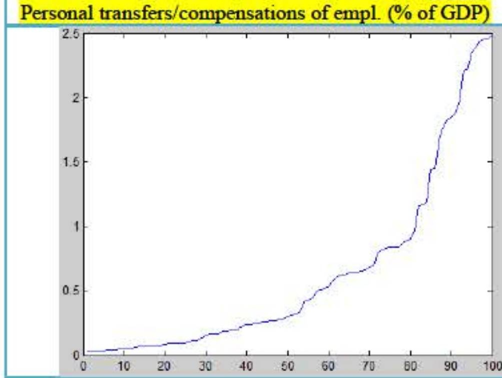
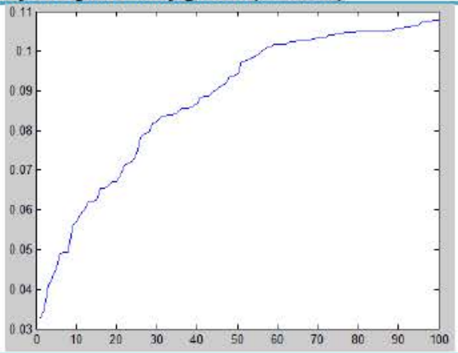
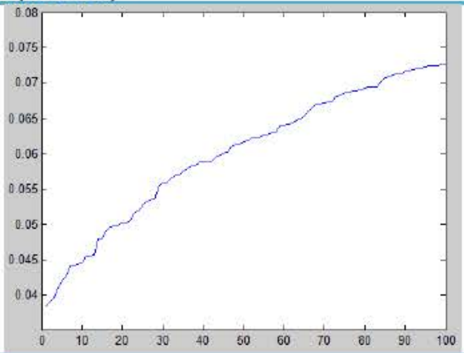
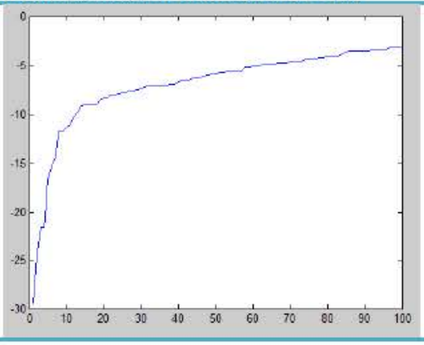
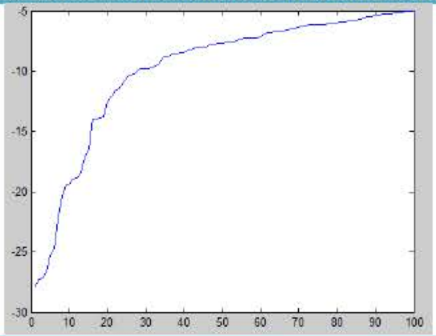
Natural Gas Rents: Top 20 Causally-Linked Series (Dynamic Time Warping)		
Raw Data	**Scaled**	**Scaled and De-Trended**
Arable land (hectares per person)	Total natural resources rents (% of GDP)	Total natural resources rents (% of GDP)
CO2 emissions (kg per 2000 US$ of GDP)	Electricity production from oil sources (% of total)	Adjusted savings: energy depletion (% of GNI)
CO2 emissions (kg per 2005 PPP $ of GDP)	Adjusted savings: natural resources depletion (% of GNI)	Adjusted savings: natural resources depletion (% of GNI)
Adjusted savings: carbon dioxide damage (% of GNI)	Electricity production from oil sources (kWh)	CO2 emissions from gaseous fuel consumption (kt)
Forest rents (% of GDP)	Adjusted savings: energy depletion (% of GNI)	CO2 emissions (kg per 2000 US$ of GDP)
CO2 emissions (kg per PPP $ of GDP)	Inflation, consumer prices (annual %)	Population in urban agglomerations of more than 1M (% total pop)
Coal rents (% of GDP)	Real interest rate (%)	Forest rents (% of GDP)
Mineral rents (% of GDP)	Inflation, GDP deflator (annual %)	Electricity production from oil sources (% of total)
CO2 emissions from other sectors (% total fuel combustion)	Total reserves in months of imports	CO2 emissions from liquid fuel consumption (% of total)
Adjusted savings: mineral depletion (% of GNI)	Oil rents (% of GDP)	CO2 emissions from residential bldgs/comm. + pub services (MMT)
Prevalence of HIV, total (% of population ages 15-49)	Adjusted net savings, excl particulate emission damage ($)	Fossil fuel energy consumption (% of total)
Permanent cropland (% of land area)	Mineral rents (% of GDP)	CO2 emissions from liquid fuel consumption (kt)
Adjusted savings: particulate emission damage (% of GNI)	Claims on private sector (annl growth as % of broad money)	Electricity production from oil sources (kWh)
Oil rents (% of GDP)	Adjusted savings: net national savings (% of GNI)	Broad money to total reserves ratio
Foreign direct investment, net inflows (% of GDP)	Adjusted savings: mineral depletion (% of GNI)	Money and quasi money (M2) to total reserves ratio
DEC alternative conversion factor (LCU per US$)	Adjusted savings: net national savings (current US$)	Quasi-liquid liabilities (% of GDP)
Official exchange rate (LCU per US$, period average)	Lending interest rate (%)	Electricity production from oil, gas and coal sources (% of total)
Foreign direct investment, net outflows (% of GDP)	Coal rents (% of GDP)	CO2 emissions from other sectors (MMT)
Armed forces personnel (% of total labor force)	Broad money growth (annual %)	Industry, value added (% of GDP)
Personal transfers/compensations of empl. (% of GDP)	Money and quasi money growth (annual %)	Oil rents (% of GDP)
		

Table 2: Automatic ranking of likely causal relationships between time series, using Granger causality

Natural Gas Rents: Top 20 Causally-Linked Series (Granger Causality)		
Raw Data	**Scaled**	**Scaled and De-Trended**
	Road sector gasoline fuel consumption per capita (kg of oil eq)	Road sector energy consumption per capita (kg of oil eq)
	CO2 emissions (metric tons per capita)	Energy imports, net (% of energy use)
	Electricity production from oil sources (kWh)	Road sector energy consumption (kt of oil equivalent)
	Energy use (kg of oil equivalent per capita)	Road sector gasoline fuel consumption (kt of oil equivalent)
	Terms of trade adjustment (constant LCU)	Road sector gasoline fuel consumption per capita (kg of oil)
	Electricity production from oil sources (% of total)	CO2 emissions from liquid fuel consumption (kt)
	Combustible renewables and waste (% of total energy)	Electricity production from oil sources (kWh)
	CO2 emissions from liquid fuel consumption (% of total)	Energy use (kt of oil equivalent)
	CO2 emissions from liquid fuel consumption (kt)	CO2 emissions from transport (million metric tons)
	CO2 emissions from solid fuel consumption (% of total)	Energy use (kg of oil equivalent per capita)
	CO2 emissions from residential buildings and commercial)	CO2 emissions (metric tons per capita)
	Adjusted net savings, ex particulate emission damage	Consumer price index (2005 = 100)
	Coal rents (% of GDP)	CO2 emissions (kt)
	CO2 emissions from manufacturing industries	Electricity production from oil sources (% of total)
	Adjusted savings: net national savings (% of GNI)	Stocks traded, turnover ratio (%)
	Adjusted savings: consumption of fixed capital (% of GNI)	Industry, value added (constant LCU)
	Land area (sq. km)	Industry, value added (constant 2000 US$)
	Surface area (sq. km)	CO2 emissions from liquid fuel consumption (% of total)
	Risk premium on lending (lending minus treasury rate)	CO2 emissions (kg per PPP $ of GDP)
	Adjusted savings: energy depletion (current US$)	Personal transfers and compensation of employees (US$)
		

Real-World Causality Analysis Use Case Development and Demonstration

We developed an in-depth, real-world demonstration use case of MAT as a causal-analysis and modeling tool, which we included in our AHFE paper and which we hope will provide a sound basis for ongoing evaluations and demonstrations. In the case study, we demonstrate a representative exploration of the causal/predictive relationship between poverty and conflict. A large body of literature exists that explores the "conflict trap"—the process whereby countries get stuck in a repeated pattern of violent conflict and economic underdevelopment (Collier et al., 2003). There have been several studies evaluating the causal/predictive link between these two features using standard statistical approaches, with some finding evidence for poverty driving societies into conflict (Collier & Hoeffler 2004, Braithwaite 2014), while others (Djankov 2008) indicate that civil conflict may be the cause of depressed economic growth. Using the methods described in the previous section, we can better untangle and characterize this relationship and gain insight into the processes that lead to the conflict trap.

The choice of data is itself a challenge for causal/predictive analyses, as the complex and abstract concepts of "poverty" and "conflict" are difficult to represent as measurable variables. To measure conflict we use the UCDP/PRIO dataset (Themnér & Wallensteen 2014), which tracks the incidence and intensity of global armed conflict between 1946 and 2013. To capture the notion of poverty, which is not merely a measure of income, but also of relative well-being, we use two variables from the World Bank World Development Indicators dataset (The World Bank 2013)—infant mortality rate, measured as the number of infants per thousand live births that die each year, and GDP, to measure the overall level of development. We consider conflict as both a categorical variable ranging from 0 to 3 indicating the intensity of a conflict in a given year, and as a numerical value with counts of the battle deaths due to conflict within a country. We focused on the timeframe from 1960-2013 as both data sets were more complete for this time period.

In our first experiment, we analyzed the relationship between poverty and conflict using Granger causality, varying the time lag between 1 and 10 years. Out of the 100 countries under study, we found strong evidence that conflict causes poverty in about 30% of the cases with a time lag of 1 year, as shown in Figure 3-9, with strength of the causal linkage degrading slightly as the time lag increased. Interestingly, there is also strong evidence of a causal relationship from poverty to conflict, but this actually consistently *increases* as we stretch out the time lag. This result may indicate the nature of conflict and poverty as persistent conditions with longer duration impacts, but may also be due to uneven time lags that cannot adequately be captured by Granger causality.

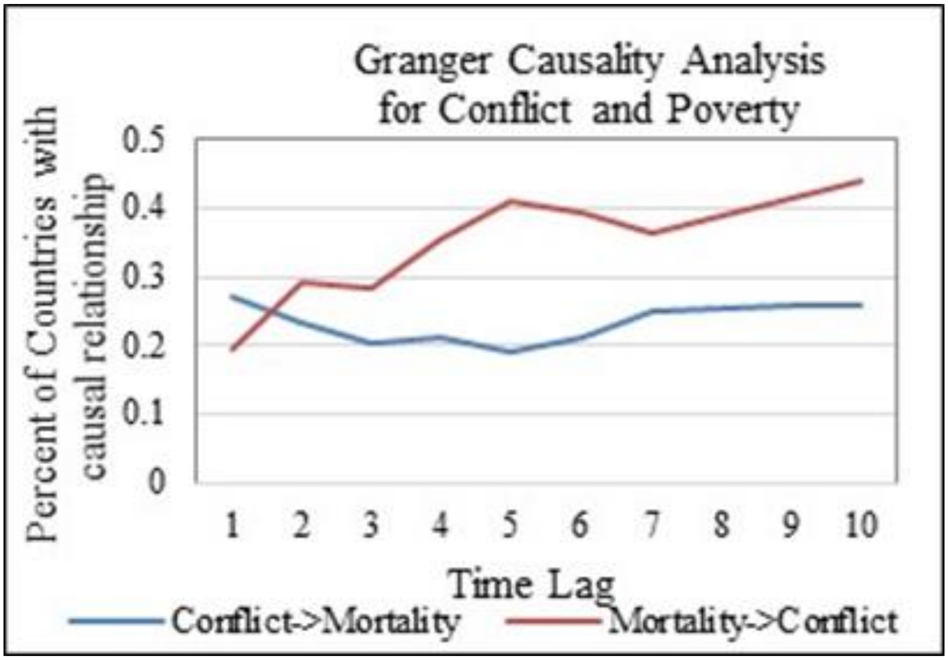


Figure 3-9. Results from Granger causality analysis with increasing time lag.

For our second experiment, we used convergent cross mapping (CCM) to further characterize the causal relationship between conflict and poverty. Social processes are often best described by complex dynamical systems, with multiple layers of feedback and interaction, and CCM can help identify these more complex causal interactions, particularly reciprocal or bidirectional causality. Because CCM examines the relationships between projections of the time series, we normalized the data to measure the percent change at each time point to account for the vastly different scales of conflict casualties, infant mortality, and GDP. We observed convergence in 80% of countries supporting the hypothesis that conflict causes poverty, and 12% for poverty leading to conflict. However, these results may be skewed by the perfect predictability of conflict in countries that experienced no conflict during the time period under study. Figure 3-10 shows an example of convergence to support the hypothesis that conflict causes poverty in Comoros.

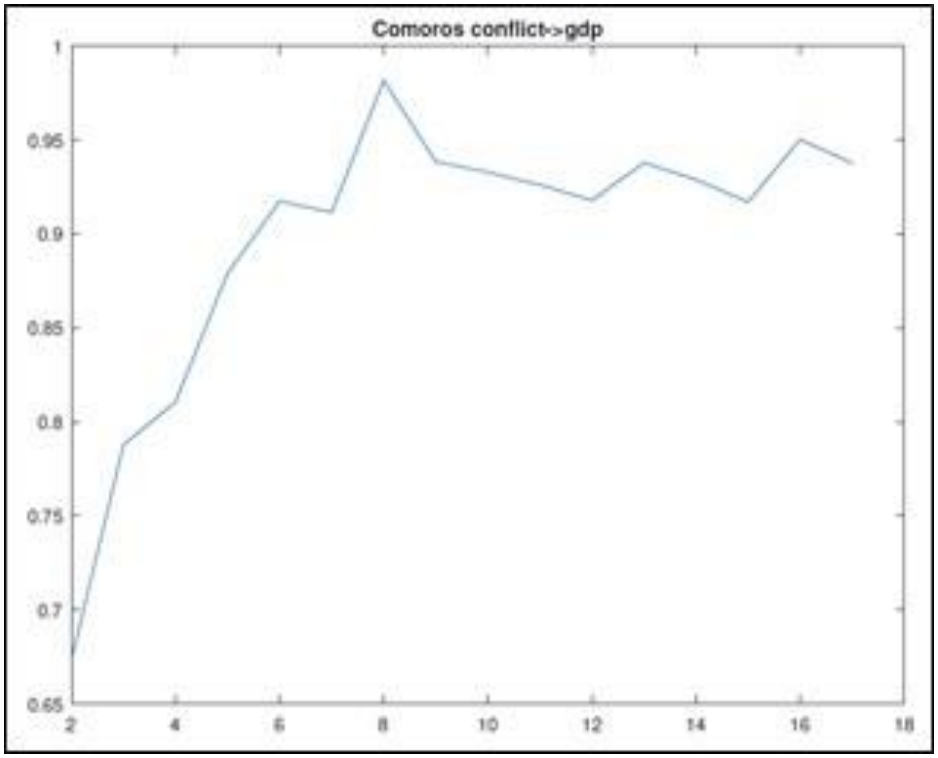


Figure 3-10. Results from CCM analysis illustrating convergence indicative of a causal link from conflict to poverty.

While dynamic time warping and convergent cross-mapping can be useful analytic tools, the nature of our case study data is not ideal for these types of methods that look explicitly for point-by-point relationships across the time series. However, even though the relationships between the conflict and poverty data are difficult to quantify through these types of measurements, we found they can be reasonably described through qualitative featurization analysis. While conflict and poverty are linked to one another, this phenomenon does not manifest as similar patterns of proportional increases or decreases in values offset in time. Instead, across the countries studied, we saw that rapid increases in conflict or periods of recurring conflict are associated not with similar fluctuations in poverty, but by continually decreasing or statically depressed levels of economic activity and by statically high levels of infant mortality. Similarly, we found that when the conflict ended, we saw decreases in poverty follow. In essence, this illustrates the notion of the conflict and poverty traps, where violence is associated not with rapid declines into poverty, but with sustained levels of minimal development.

Figure 3-11 shows an example using the qualitative feature-based approach to analyze the data from Senegal. The top plot indicates GDP in current US$ from 1989 to 2013, the middle plot shows the number of battle related deaths, and the bottom chart is the infant mortality rate. The human-guided qualitative featurization algorithm has divided these data series into important component pieces representing distinct features. From these features, it is evident that there was a period of violent conflict from 1989 to about 2004, with several spikes in the number of casualties. During this same time, infant mortality was consistently high, and GDP was consistently low. However, after 2004, GDP and infant mortality both begin to steadily improve, while conflict remains very limited. Using these features to represent concepts such as “spikes in conflict” and “high infant mortality,” we can identify causal patterns between these more complex features that are not visible when doing a lower-level comparison of individual time points.

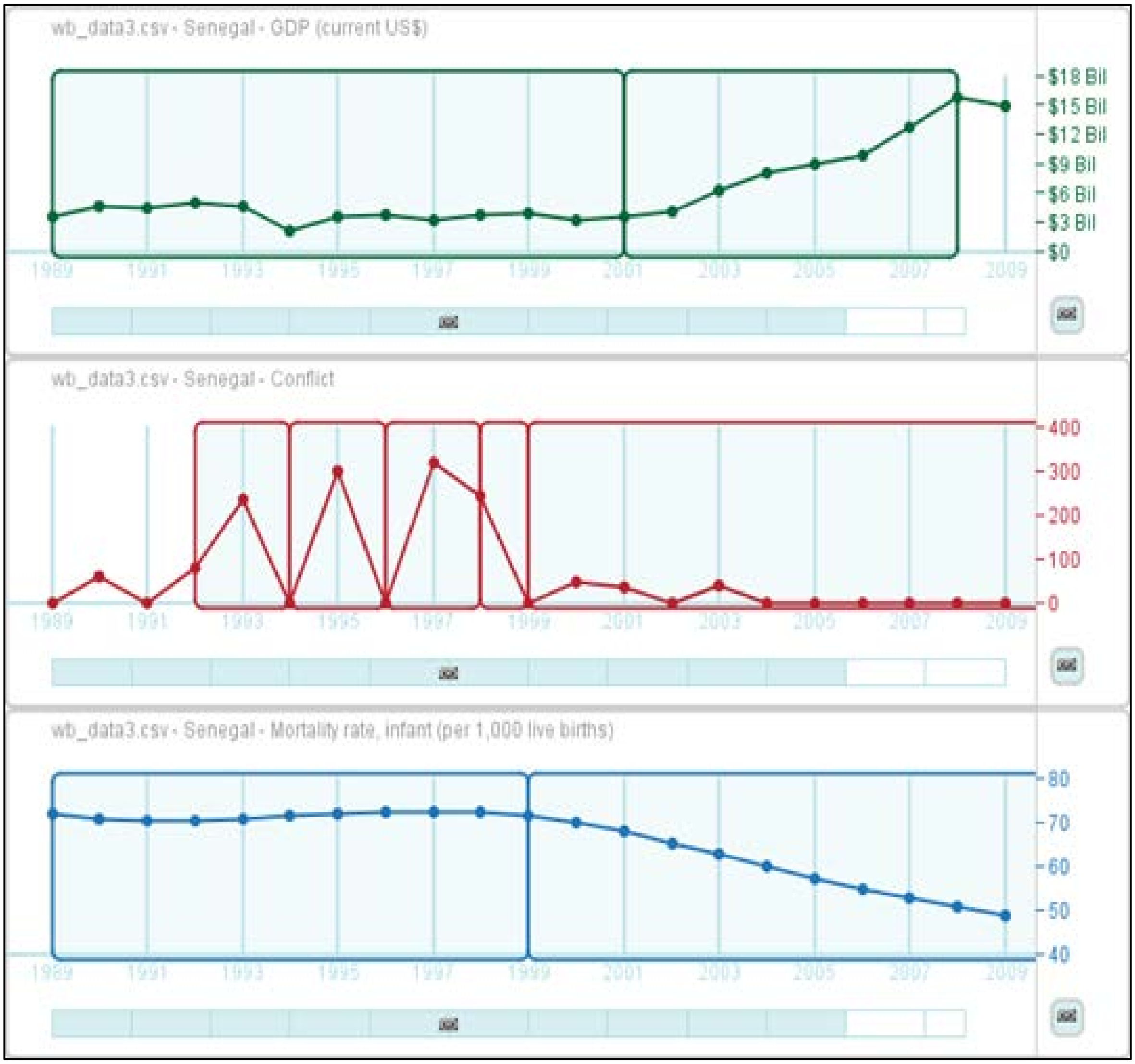


Figure 3-11. Qualitative featurization showing the relationships betcapween poverty and conflict.

3.2. Automated Causal Model Recommendations

3.2.1. Automated Suggestions for Model Construction and Refinement

The Model Construction and Refinement capability in MAT is used to suggest changes to the causal model to improve its explanatory power. We developed a model recommendation tool that can discover feature series with potential explanatory value and suggest possible explanations to the user. The design concept for this tool is that if users discover that there are features of the data that are not well explained by their theory, the model recommendation tool can help find alternative explanations that can be used to refine their theory.

In this section, we first describe the process for automatically identifying and recommending model refinements, then we describe a new capability for automatically identifying events in the data that have not been explicitly identified by the user but that might still have explanatory power.

3.2.2. Causal Model Recommendations

The causal model recommender automatically suggests modifications to user-defined causal models from the available data. MAT attempts to return models that are optimal along the following dimension:

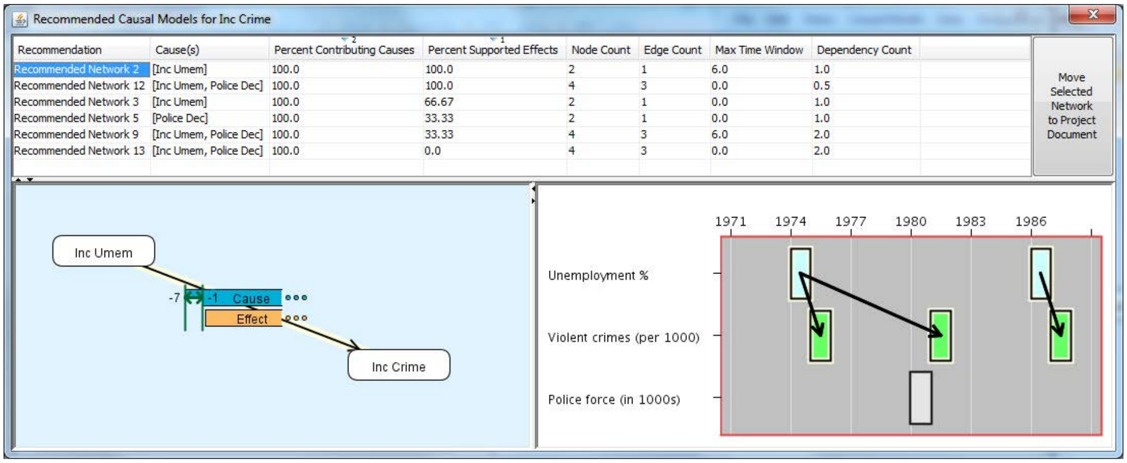
- Performance [The number of supported effects and contributing causes]. MAT attempts to find models that explain more events, both in terms of not having causes without corresponding effects and not having effects with corresponding causes.
- Model size [The number of nodes and edges]. MAT prefers simpler models over more complex models. This is both an Occam's-Razor like heuristic, but also an attempt to not overfit data, which is a problem in many machine learning systems that are able to explain all data but only through the creation of overly complex and, more importantly, typically wrong models.
- Temporal aspects [The duration of temporal windows]. MAT looks for models that use temporally closer causes and effects. That is a cause early in the dataset could be used to explain all events from then forward if the duration of the causes impact is assumed to be arbitrarily long. Most causes, however, tend to be temporally close to their effects.

MAT returns a set of suggested model edits and display the results in a user-friendly way. The recommendations are now a list of causal models where no model is strictly worse that another model in the list (that is, worse along all three dimensions just described). This set of models are said to define the Pareto Frontier, a term from multi-objective optimization to describe this kind of solution set for problems with different tradeoffs in terms of the objectives they are attempting to satisfy. This approach eliminates any obviously poor recommendations, but also makes no assumptions about the user's preferences regarding tradeoffs between various characteristics of the causal models.

The recommendations include simple causal models that only have a single cause for the effect of interest, but more complex causal models are also generated where multiple causes are combined using logic nodes. We chose to make only single-cause model recommendations because we believed these would be the most plausible and acceptable modifications to the model created by the user.

MAT currently uses two algorithms for generating causal model recommendations. The first examines all possible combinations of causes with all possible combinations of temporal offsets for the data provided. This approach quickly becomes computationally expensive, so a second approach is also included with more complex models built using the results from simpler models, resulting in the reduction of the number of possible causal models to evaluate. However, this approach may miss a causal model (e.g., models with multiple causes) that is found by the first approach. The recommender displays a progress bar during the operation and the user can cancel the operation if it is taking too long or if a model is found that seems acceptable or interesting.

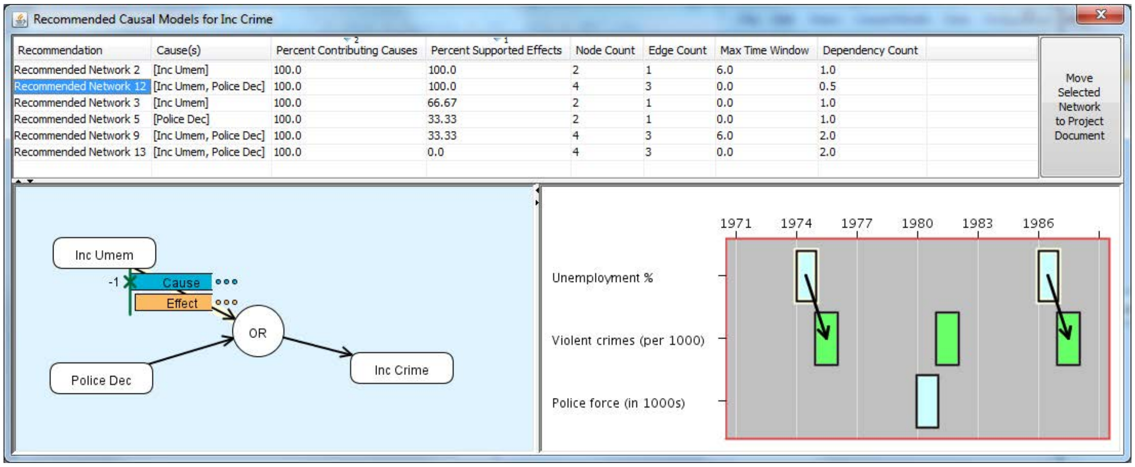
Both of these algorithms return a Pareto Frontier of causal models, which prevents any obviously inferior models from being presented to the user, but many recommendations may still be generated. Therefore, the recommender results are displayed in a sortable table. Each row is a causal model recommendation and the user can sort based on various aspects of the causal models by clicking on the column headers in the table. The recommendation table makes it easy for the user to explore the various causal models and to see how it influences model validation. Figure 3-12 shows the user looking at a simple causal model with a large temporal window. That is the model requires the effects of the cause to affect events up to 7 years in the future (see the “-7” in the annotation on the link in the figure). Figure 3-13 shows a more complex causal model (there are multiple disjunctive possible causes), but with a smaller temporal window (causes only need to have an effect of 1 year).



Recommended Causal Models for Inc Crime

Recommendation	Cause(s)	Percent Contributing Causes	Percent Supported Effects	Node Count	Edge Count	Max Time Window	Dependency Count
Recommended Network 2	[Inc Umem]	100.0	100.0	2	1	6.0	1.0
Recommended Network 12	[Inc Umem, Police Dec]	100.0	100.0	4	3	0.0	0.5
Recommended Network 3	[Inc Umem]	100.0	66.67	2	1	0.0	1.0
Recommended Network 5	[Police Dec]	100.0	33.33	2	1	0.0	1.0
Recommended Network 9	[Inc Umem, Police Dec]	100.0	33.33	4	3	6.0	2.0
Recommended Network 13	[Inc Umem, Police Dec]	100.0	0.0	4	3	0.0	2.0

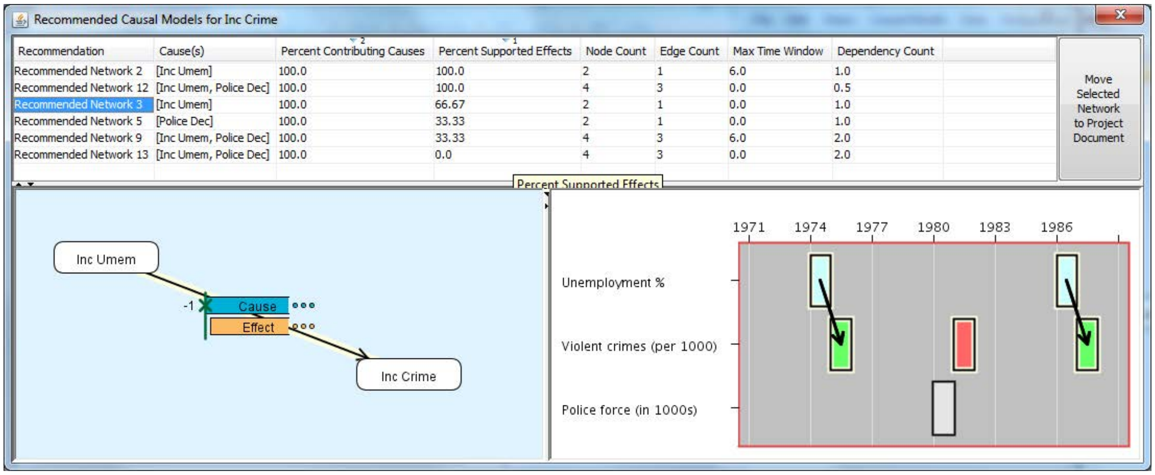
Figure 3-12: Simple causal model


Recommended Causal Models for Inc Crime
Move Selected Network to Project Document
Inc Umem
-1
Cause
Effect
OR
Police Dec
Inc Crime
1971
1974
1977
1980
1983
1986
Unemployment %
Violent crimes (per 1000)
Police force (in 1000s)

Recommendation	Cause(s)	Percent Contributing Causes	Percent Supported Effects	Node Count	Edge Count	Max Time Window	Dependency Count
Recommended Network 2	[Inc Umem]	100.0	100.0	2	1	6.0	1.0
Recommended Network 12	[Inc Umem, Police Dec]	100.0	100.0	4	3	0.0	0.5
Recommended Network 3	[Inc Umem]	100.0	66.67	2	1	0.0	1.0
Recommended Network 5	[Police Dec]	100.0	33.33	2	1	0.0	1.0
Recommended Network 9	[Inc Umem, Police Dec]	100.0	33.33	4	3	6.0	2.0
Recommended Network 13	[Inc Umem, Police Dec]	100.0	0.0	4	3	0.0	2.0

Figure 3-13: More complex causal model

Causal models that have lower performance, but excel in other aspects are also included in the table. Figure 3-14 shows a simple causal model with a small temporal window. This selection may be preferred to either of the previous two models even though it does not provide support for all of the effects.


Recommended Causal Models for Inc Crime
Move Selected Network to Project Document
Percent Supported Effects
Inc Umem
-1
Cause
Effect
Inc Crime
1971
1974
1977
1980
1983
1986
Unemployment %
Violent crimes (per 1000)
Police force (in 1000s)

Recommendation	Cause(s)	Percent Contributing Causes	Percent Supported Effects	Node Count	Edge Count	Max Time Window	Dependency Count
Recommended Network 2	[Inc Umem]	100.0	100.0	2	1	6.0	1.0
Recommended Network 12	[Inc Umem, Police Dec]	100.0	100.0	4	3	0.0	0.5
Recommended Network 3	[Inc Umem]	100.0	66.67	2	1	0.0	1.0
Recommended Network 5	[Police Dec]	100.0	33.33	2	1	0.0	1.0
Recommended Network 9	[Inc Umem, Police Dec]	100.0	33.33	4	3	6.0	2.0
Recommended Network 13	[Inc Umem, Police Dec]	100.0	0.0	4	3	0.0	2.0

Figure 3-14: Another causal model with different properties/tradeoffs

3.2.3. Automatic Feature Extraction and Pattern Analysis

In many domains, causal models can be most easily described in terms of patterns of qualitative features, rather than quantitative relationships. In MAT, users can identify qualitative features in data streams that represent meaningful events, such as “spikes in crime.” The existing feature recognition system uses these user-identified events as exemplars in a learning-by-example approach, automatically searching for repeated, temporal patterns of these events in the data.

This only works, however, when the user knows which features are of interest ahead of time. We expect this often be the case, but not always, so we included functionality in MAT to automatically mine the available data for "interesting" features that have explanatory power for causes of other user-defined events. To provide this capability, we developed an automated approach to extracting features in data streams using a non-linear optimization algorithm, the Nelder-Mead Simplex algorithm, to identify structural, qualitative features of a data series. This algorithm divides a time series into the optimal combination of structural features using the featurization "language" (from Olszewski, 2001) discussed in previous reports (see Figure 3-15).

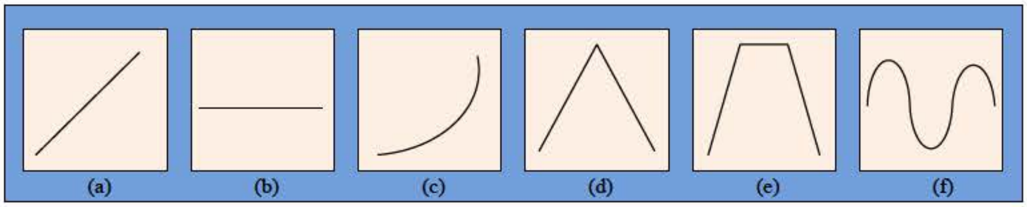
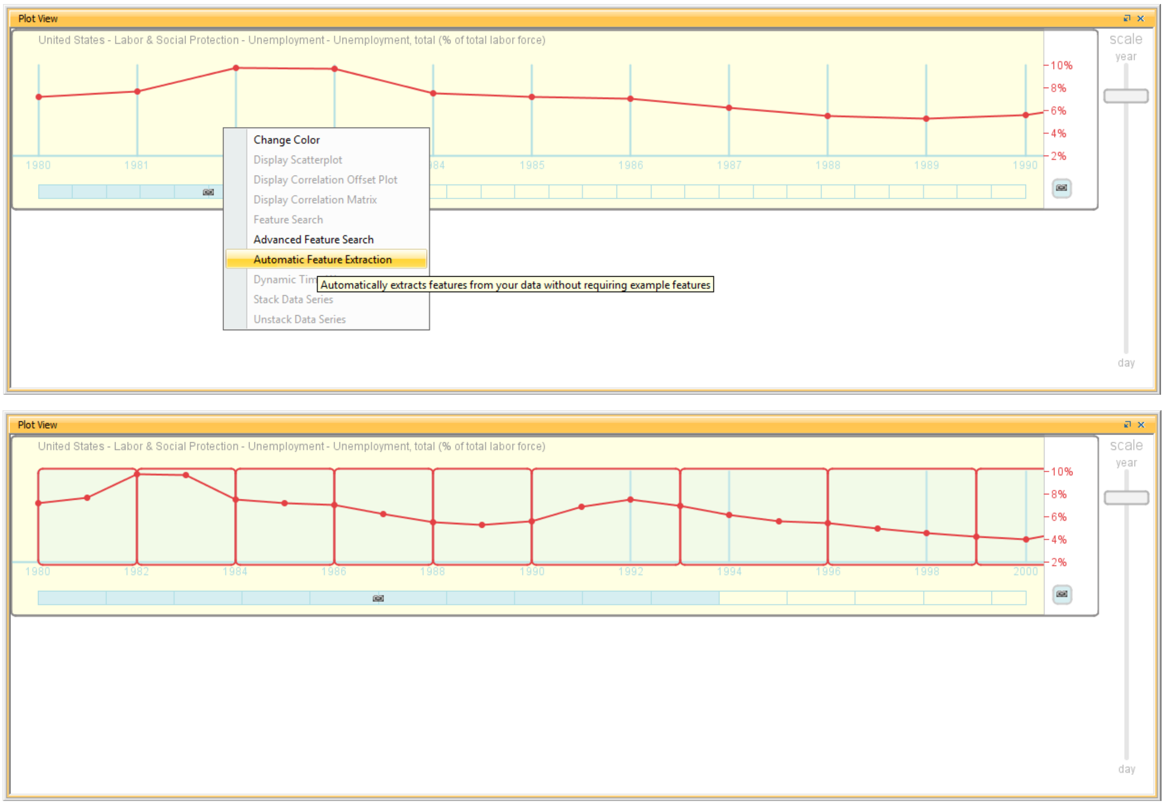


Figure 3-15: Six common function morphologies that can comprise qualitative features: (a) slope, (b) constant, (c) exponential, (d) triangle, (e) trapezoidal, (f) sinusoidal

When features are selected by the automatic feature extraction algorithm, they are then clustered into *meaningful* concepts. For example, similarly shaped exponential increases in crime are grouped together in a concept called "increases in crime." This mechanism is based on the morphologies shown in Figure 3-15, but we plan to explore additional clustering algorithms that can group features at a finer granularity according to the parameters of their structural representation and their duration over time. This new capability has been fully integrated into the MAT user interface (see Figure 3-16).


Plot View
United States - Labor & Social Protection - Unemployment - Unemployment, total (% of total labor force)
10%
8%
6%
4%
2%
1980
1981
1985
1986
1987
1988
1989
1990
scale
year
day
Change Color
Display Scatterplot
Display Correlation Offset Plot
Display Correlation Matrix
Feature Search
Advanced Feature Search
Automatic Feature Extraction
Dynamic Tim
Stack Data Series
Unstack Data Series
Plot View
United States - Labor & Social Protection - Unemployment - Unemployment, total (% of total labor force)
10%
8%
6%
4%
2%
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
scale
year
day

Automatically extracts features from your data without requiring example features

Figure 3-16: Automatic feature extraction identifies qualitative structures in a time series

This approach fully featurizes a data stream, which can generate more features that are useful or interesting. We combined this automatic feature extraction with a heuristic version of the TF-IDF (term frequency-inverse document frequency) algorithm from document analysis to identify features that are characteristic of a time series (i.e., frequent in the data stream, but infrequent in other data streams), and those that are uncommon, but extreme and meaningful from a causal modeling perspective (e.g., you might only have one stock market crash in your data, so it is not frequent, but is extreme enough to be interesting).

In MAT, this automated feature extraction can be used with the causal model recommender, providing additional candidate causes. With this capability, MAT now provides the user with novel suggestions of causal relationships based on features that might otherwise have been overlooked, helping user refine and validate their causal models.

3.3. Data Validation

Data validation will support data analysts who are concerned that they are developing models with unreliable data or who are developing models to help detect when a system breaks. For instance, if we can develop a model for how survey data results from various regions tend to correspond to each other, then we can develop tools that can flag data that appears to violate those historic patterns, and so might not be appropriate for use in decision making or model building.

The basic idea for our approach is similar to that described in Dereszynski and Dietterich (2007, 2011) in the domain of environmental science. That is, we create a graphical, probabilistic model from "good" data that models and learns how various data sources relate to each other. For MAT, we plan to have the use identify the structure (that is, which variable are likely to related) and to use machine learning techniques to learn the patterns (traditional multi-linear regression would be acceptable in many cases, but we also happen to have access to richer probabilistic machine learning tools in house that let us build more sophisticated models if needed). We use the resulting model to identify (and suggest repairs for) anomalies in a dataset that is not known ahead of time to be "good." We currently assume that a domain expert user can recognize "good" data from which to learn.

We have identified two basic reusable templates to describe data gathering situation, the first is where there is a sensible controller for each variable of interest, and the second is where there is no controller. An example of the first case is a simpler thermostat where you can sense the desired temperature value as specified by the thermostat and you can sense the actual temperature. In this case, either the controller or the sensor can be broken and the task is to identify which and to estimate the true value based on the available data. For instance, if the sensed temperature begins returning 0 over and over or a random value, we might guess that they temperature sensor is broken and estimate that the true value is actually the thermostat value.

An example of the second case is where we have polling data from a number of villages within the same region. In general, they will tend to vary a bit from each other, but will also be highely correlated. In this case, imagine a poll-taker who decides to fill out the polls himself randomly instead of asking people, you would see a shift in that village from the others that should be detectable. In this case, we could hypothesize that the actual value is roughly the value of the other villages (or some learned offset if, say, one village regularly polls a bit higher or lower on a question).

For both of these cases we have identified a simple graphical template to describe the situation and a mapping from that template into a graphical probabilistic model to describes the situation in a bit more detail. For instance, for the thermostat example is depicted in Figure 3-17 (note that it is possible to create multiple thermostats, such as in a home with multiple heating zones).

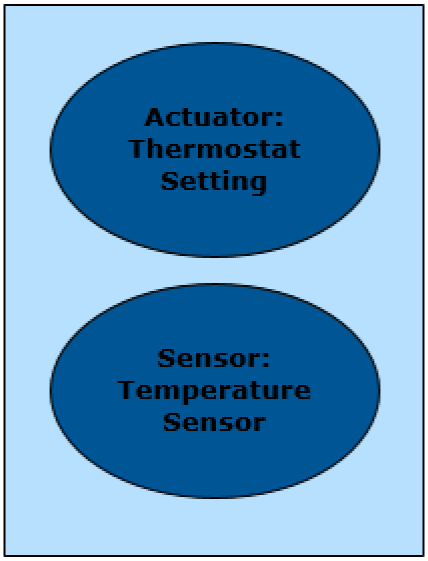


Figure 3-17. Simple thermostat scenario

In this case, the scenario maps into the graphical model shown in Figure 3-18. This figure is obviously quite a bit more complex, which is why we want the user to work in the simpler and more intuitive space demonstrated in Figure 3-17. A key reason for the complexity of the probabilistic model is that it learns how things change over time and, so, can learn that temperatures change gradually from the current temperature to a new specified temperature and that something that is broken tends to stay broken.

The probabilistic model is encoded in the open source Figaro programming language, which can take data as input and learn the model parameters that describe the behavior of the model. Once these parameters have been learned, the model can detect broken sensors and broken thermostats and can make estimates of actual values of the parameters that cannot be properly sensed. For now the Figaro code is written by hand, but is structured according to the templates. Eventually, we would want to automatically generate the Figaro code as well.

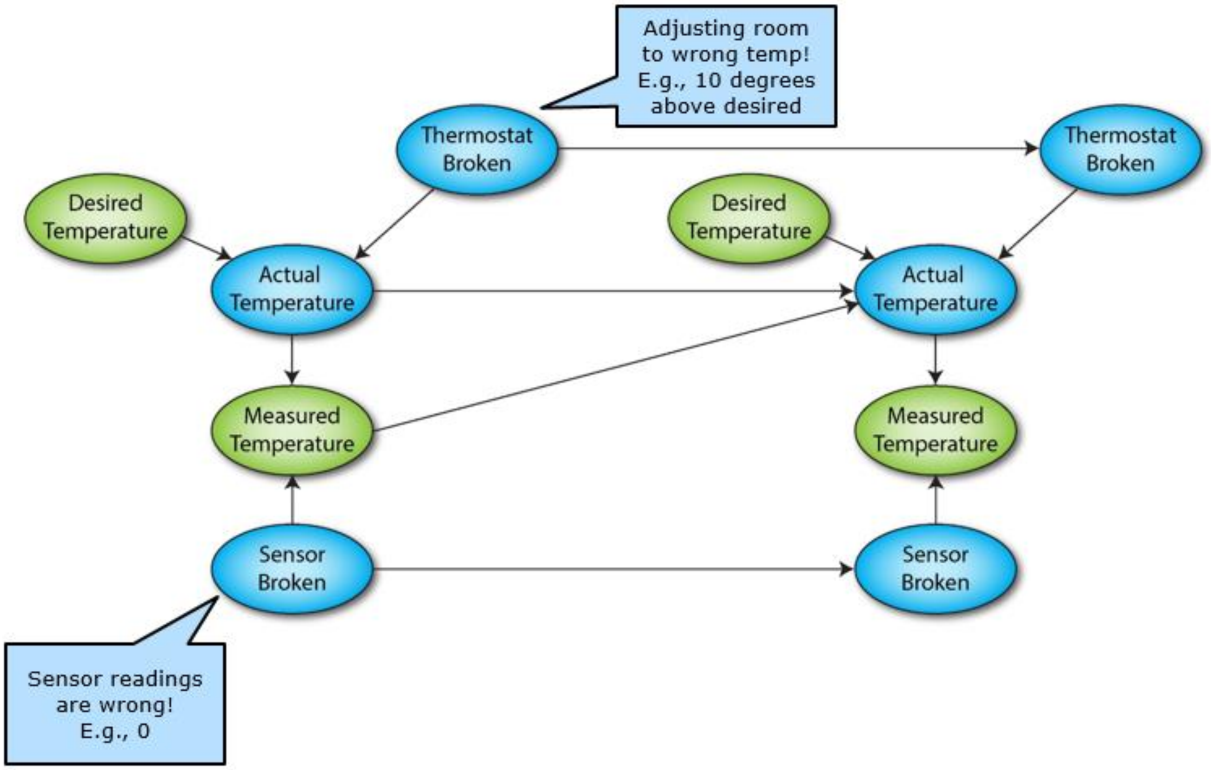


Figure 3-18. Graphical probabilistic model of thermostat scenario

Figure 3-19 shows the learning system in action. In this case, we see the thermostat being set to different values throughout the day, the temperature changing accordingly, and the systems estimates largely tracking the actuals. Figure 3-20 shows the systems underlying assessment of whether any of the data is bad. There is a blip in the middle where there is estimated to be some chance that the thermostat is broken (because the temperature is declining after turning down the thermostat more slowly than in the training data), but it never gets high enough to believe there is a significant chance that there is a problem.

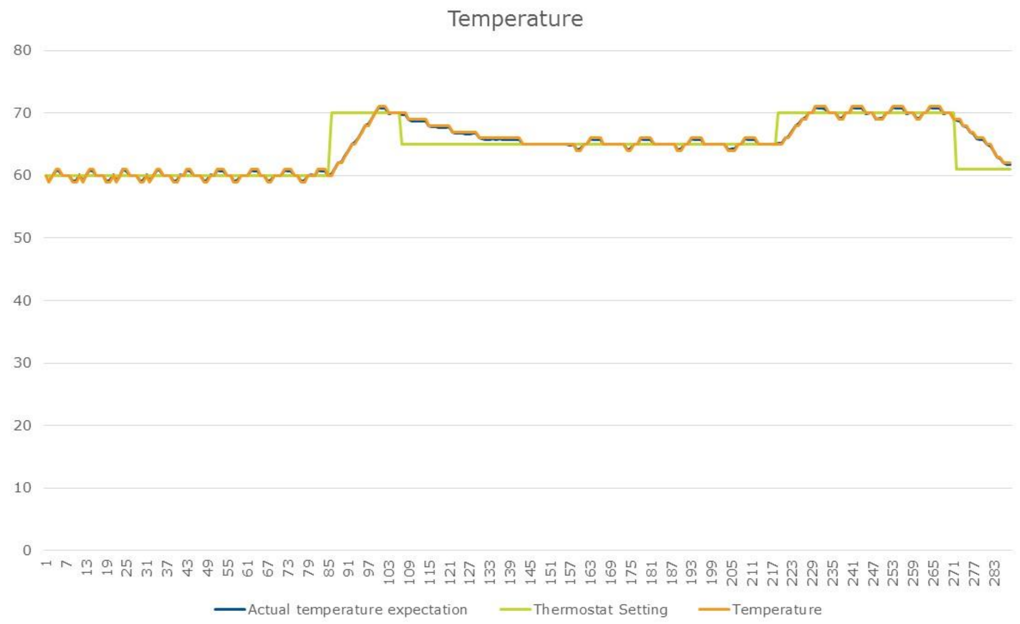


Figure 3-19. Temperature data validation in no-error case

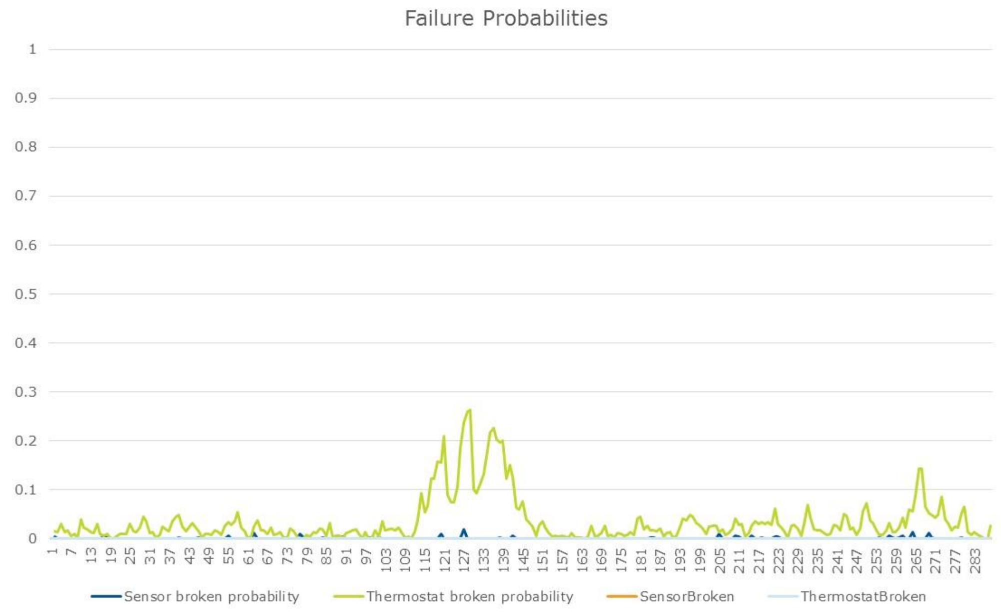


Figure 3-20. Estimating broken sensor/actuator

Figure 3-21 and Figure 3-23 present two other scenarios where there are errors to detect. In the first case, the temperature sensor is broken; in the second, the thermostat itself is broken. In both cases, we see that the system can fairly rapidly detect that the data is improbable and can make reasonable estimates of the actual values. Figure 3-22 and Figure 3-24 show the estimates of error probabilities.

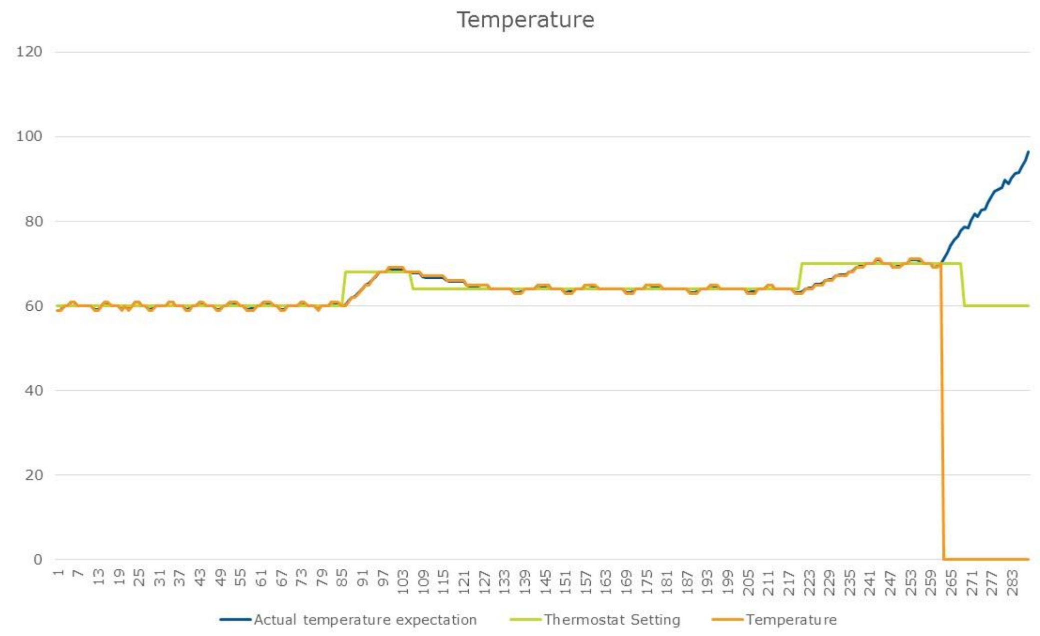


Figure 3-21. Broken temperature sensor scenario

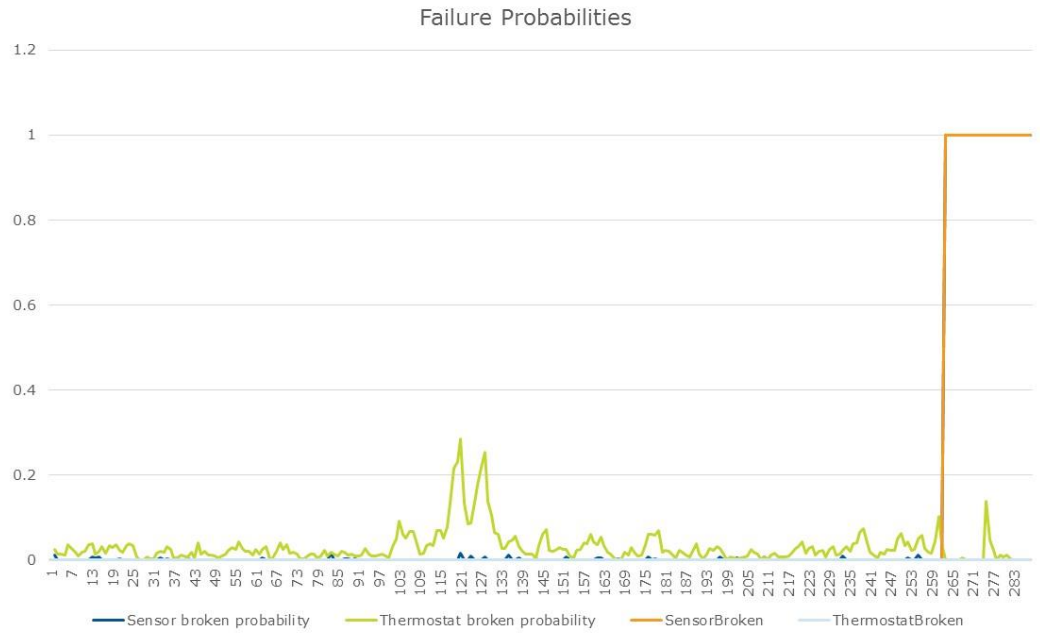


Figure 3-22. Probabilistic estimates of broken thermostat

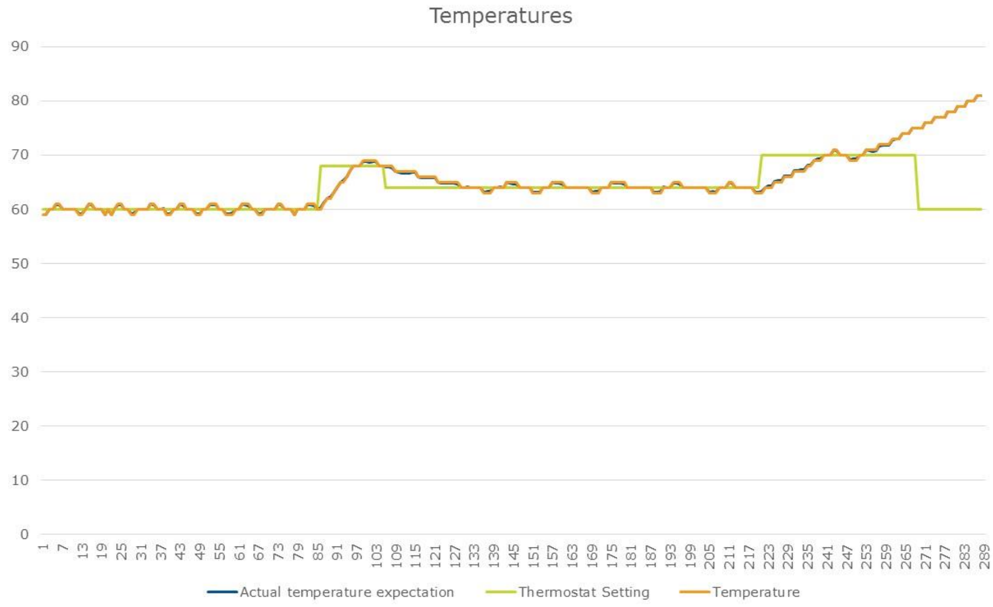


Figure 3-23. Broken thermostat scenario

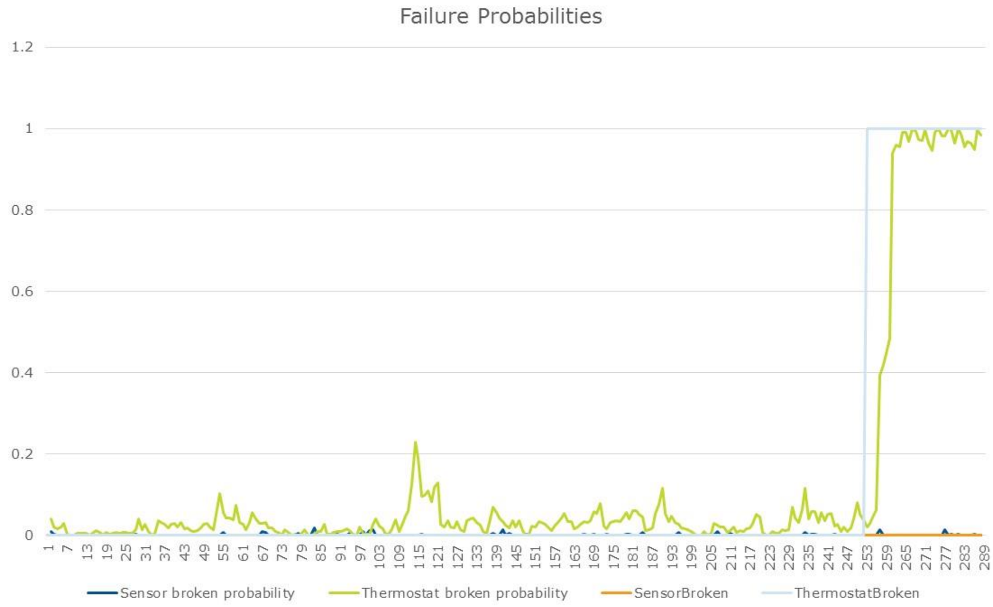


Figure 3-24. Probabilistic estimates of broken thermostat

Figure 3-25 is an example of three villages each responding to a poll. In this case, there are no arrows in the figure as the links are undirected, making it a Markov model instead of a Bayesian model. An advantage of Figaro is that is makes it simple to encode both kinds of models in the same framework. Figure 3-26 shows the graphical probabilistic model that is derived in this situation, which has a combination of undirected and directed links, again something that Figaro supports naturally. This model is actually simpler than it could be as we have made no attempt to model dynamics over time like we did in the thermostat case. This is not especially complex to do programmatically, but makes the figure much more complicated.

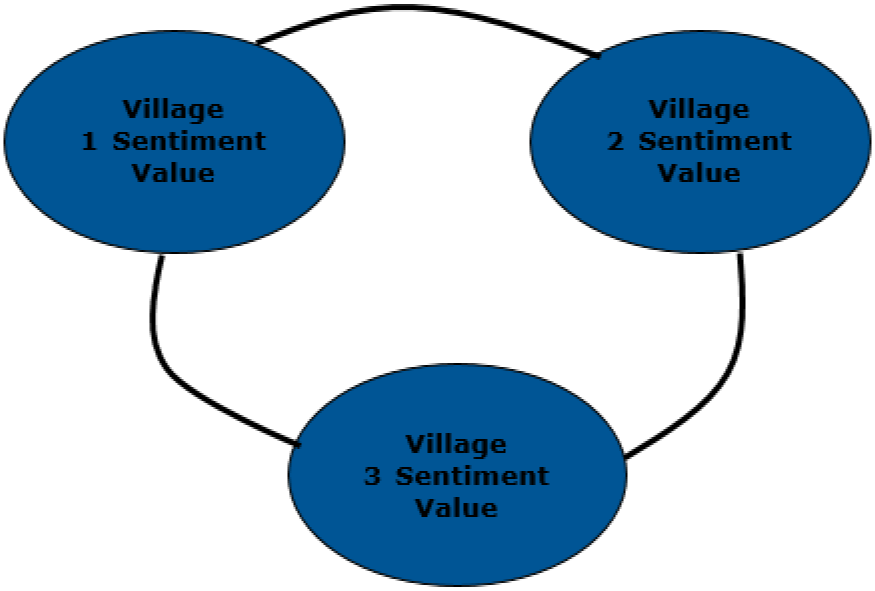

Village 1 Sentiment Value
Village 2 Sentiment Value
Village 3 Sentiment Value

Figure 3-25. Three-Village Polling Scenario

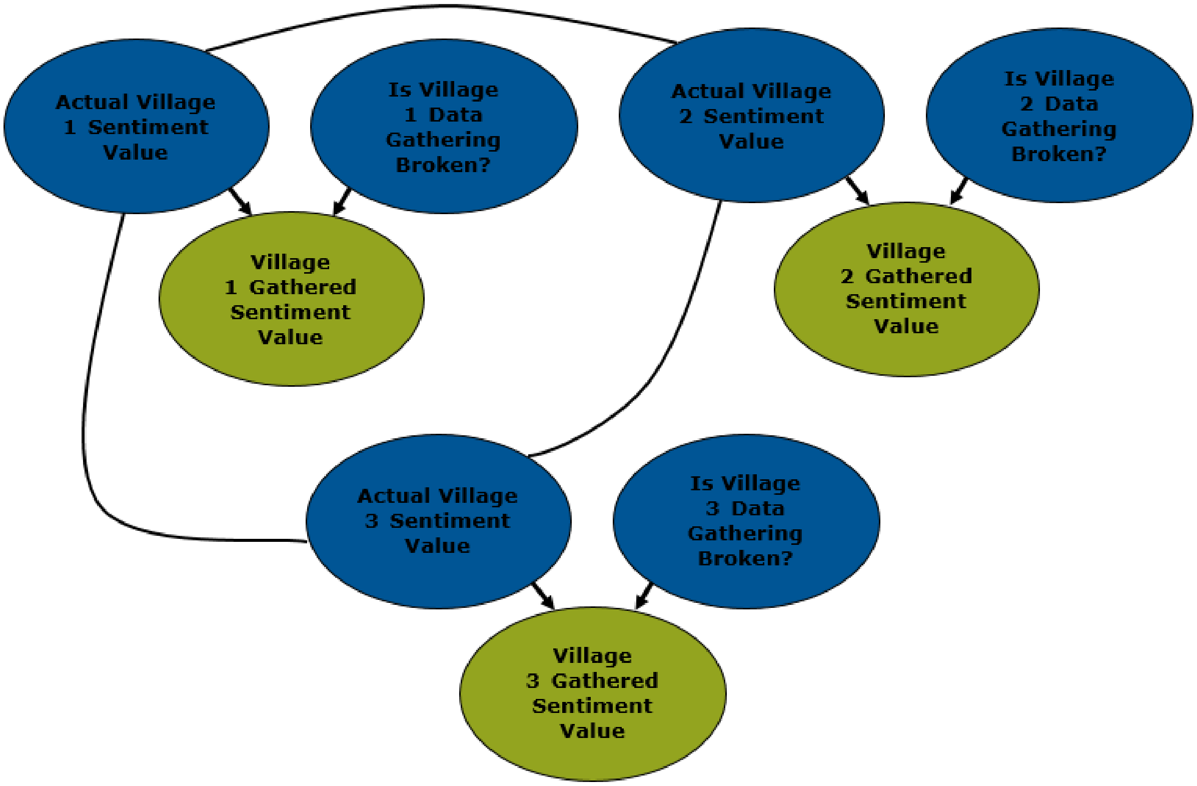
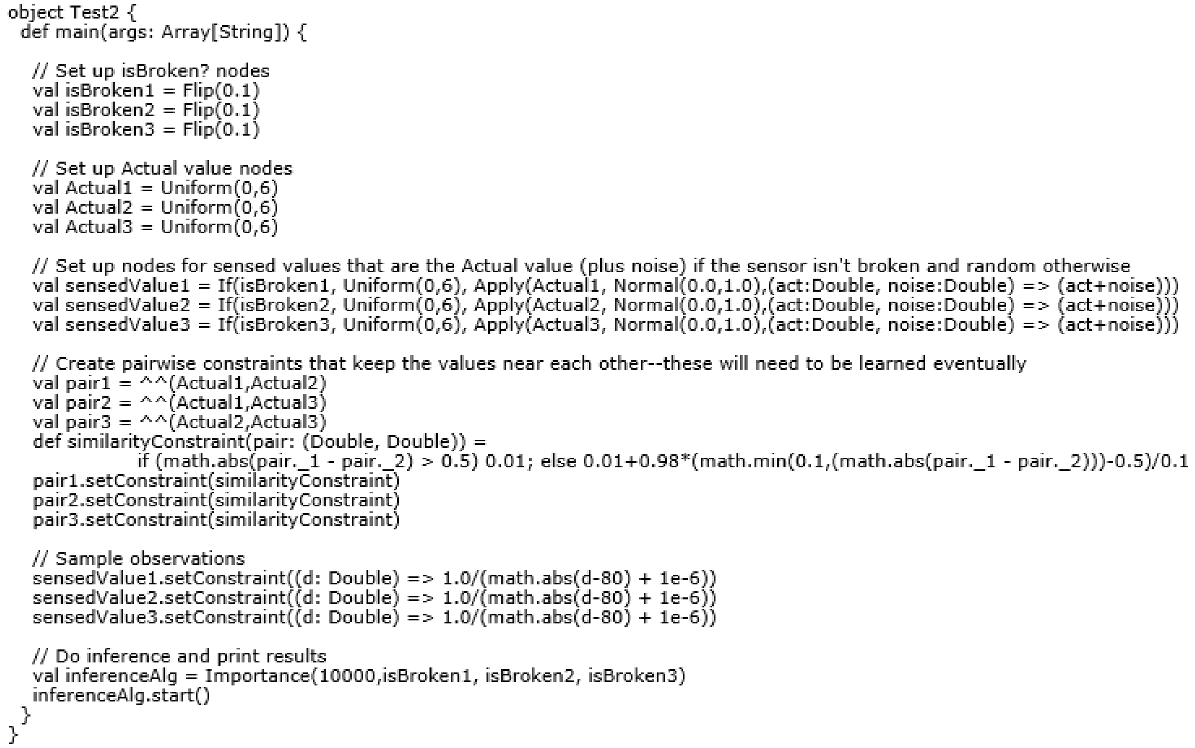

Actual Village 1 Sentiment Value
Is Village 1 Data Gathering Broken?
Actual Village 2 Sentiment Value
Is Village 2 Data Gathering Broken?
Village 1 Gathered Sentiment Value
Village 2 Gathered Sentiment Value
Actual Village 3 Sentiment Value
Is Village 3 Data Gathering Broken?
Village 3 Gathered Sentiment Value

Figure 3-26. Graphical Probabilistic Model of the Three-Village Polling Scenario

Finally, the Figaro code that describes this model is shown in Figure 3-27. As noted above, this code is still generated by hand, but this is not especially cumbersome once one knows the Figaro language, and we expect to automate even this in the future. Note that there are some hard-coded parameters in the model. These can be hard-coded if there is a domain expert to specify them, but they can also be learned using Figaro's built-in learning mechanisms.



```
object Test2 {
  def main(args: Array[String]) {

    // Set up isBroken? nodes
    val isBroken1 = Flip(0.1)
    val isBroken2 = Flip(0.1)
    val isBroken3 = Flip(0.1)

    // Set up Actual value nodes
    val Actual1 = Uniform(0,6)
    val Actual2 = Uniform(0,6)
    val Actual3 = Uniform(0,6)

    // Set up nodes for sensed values that are the Actual value (plus noise) if the sensor isn't broken and random otherwise
    val sensedValue1 = If(isBroken1, Uniform(0,6), Apply(Actual1, Normal(0.0,1.0),(act:Double, noise:Double) => (act+noise)))
    val sensedValue2 = If(isBroken2, Uniform(0,6), Apply(Actual2, Normal(0.0,1.0),(act:Double, noise:Double) => (act+noise)))
    val sensedValue3 = If(isBroken3, Uniform(0,6), Apply(Actual3, Normal(0.0,1.0),(act:Double, noise:Double) => (act+noise)))

    // Create pairwise constraints that keep the values near each other--these will need to be learned eventually
    val pair1 = ^^(Actual1,Actual2)
    val pair2 = ^^(Actual1,Actual3)
    val pair3 = ^^(Actual2,Actual3)
    def similarityConstraint(pair: (Double, Double)) =
                if (math.abs(pair._1 - pair._2) > 0.5) 0.01; else 0.01+0.98*(math.min(0.1,(math.abs(pair._1 - pair._2)))-0.5)/0.1
    pair1.setConstraint(similarityConstraint)
    pair2.setConstraint(similarityConstraint)
    pair3.setConstraint(similarityConstraint)

    // Sample observations
    sensedValue1.setConstraint((d: Double) => 1.0/(math.abs(d-80) + 1e-6))
    sensedValue2.setConstraint((d: Double) => 1.0/(math.abs(d-80) + 1e-6))
    sensedValue3.setConstraint((d: Double) => 1.0/(math.abs(d-80) + 1e-6))

    // Do inference and print results
    val inferenceAlg = Importance(10000,isBroken1, isBroken2, isBroken3)
    inferenceAlg.start()
  }
}
```

Figure 3-27. Figaro code for three-village polling scenario

3.4. Application to Multiple Scientific Domains

3.4.1. Technical Issues Involved in Application to Multiple Scientific Modeling Domains

As we applied MAT to data in new scientific domains, we found a number of extensions necessary to maximize the applicability of MAT in those domains. In particular, we added support for handling:

- Different time scales
- Different types and scales of data
- Datasets that are significantly larger than we have encountered in social science models

We describe these three efforts in the following sections.

Handling Time

As we began to apply MAT to scientific domains beyond social science, and to certain social science datasets, we found the biggest hurdles were associated with importing and visualizing data in very different time scales. To address these issues, we improved and generalized how we handle time within MAT.

MAT datasets are time series. The Java `Date` class was used to represent the time value (the x-axis). This was convenient because there are built-in methods in Java to handle this type and perform calculations on it, but it can represent a small range of possible time values. For example, Java `Dates` cannot consistently handle time values before 1970 and cannot handle virtual or relative time values, such as microseconds or nanoseconds, that might be used in physical experiments.

To expand our ability to handle a full range of possible time values, we created an extensible, unit-neutral time representation system. MAT now supports data in microseconds, nanoseconds, and femtoseconds, milliseconds, years, and provides the user the ability to add their own times scales (e.g., geologic time). Any historic time value, including B.C. values, can be handled. The system supports all algebraic operations with time amounts and conversion and comparison between different time units. Data series can now start from an arbitrary zero point, which is better suited to scientific experiments than the previous method of starting at a specific date.

Making these improvements required wide-ranging changes and enhancements that affected the whole application and virtually every section of code in the project was involved. For example, the data import window and features required update. This window was expanded to handle the new range of time data and types (such as arbitrary reference times and frequency-based time intervals).

Handling New Types and Scales of Data

One of the biggest problems we encountered when handling new types of data was how to visualize the labels for the data in a clear, readable manner. To solve this problem, we made several improvements to the Y-axis in the data visualization to make it more readable and informative. The algorithm picks "reasonable" increments approximating the actual range of the data series (for example, changes of 1, 2, 5, or 10 units), then formats them in a human-readable way for display on the Y axis. The values are then formatted in one of several ways, such as exponential format. Figure 3-28 and Figure 3-29 show examples of this new functionality. Figure 3-28 shows how MAT has converted the Y-axis labels to suitable exponential notation. Figure 3-29 shows a variety of different data scales and types, including dollars, percentages, and a variety of value ranges being handled, such as by turning "4,500,000" into "4.5 M."

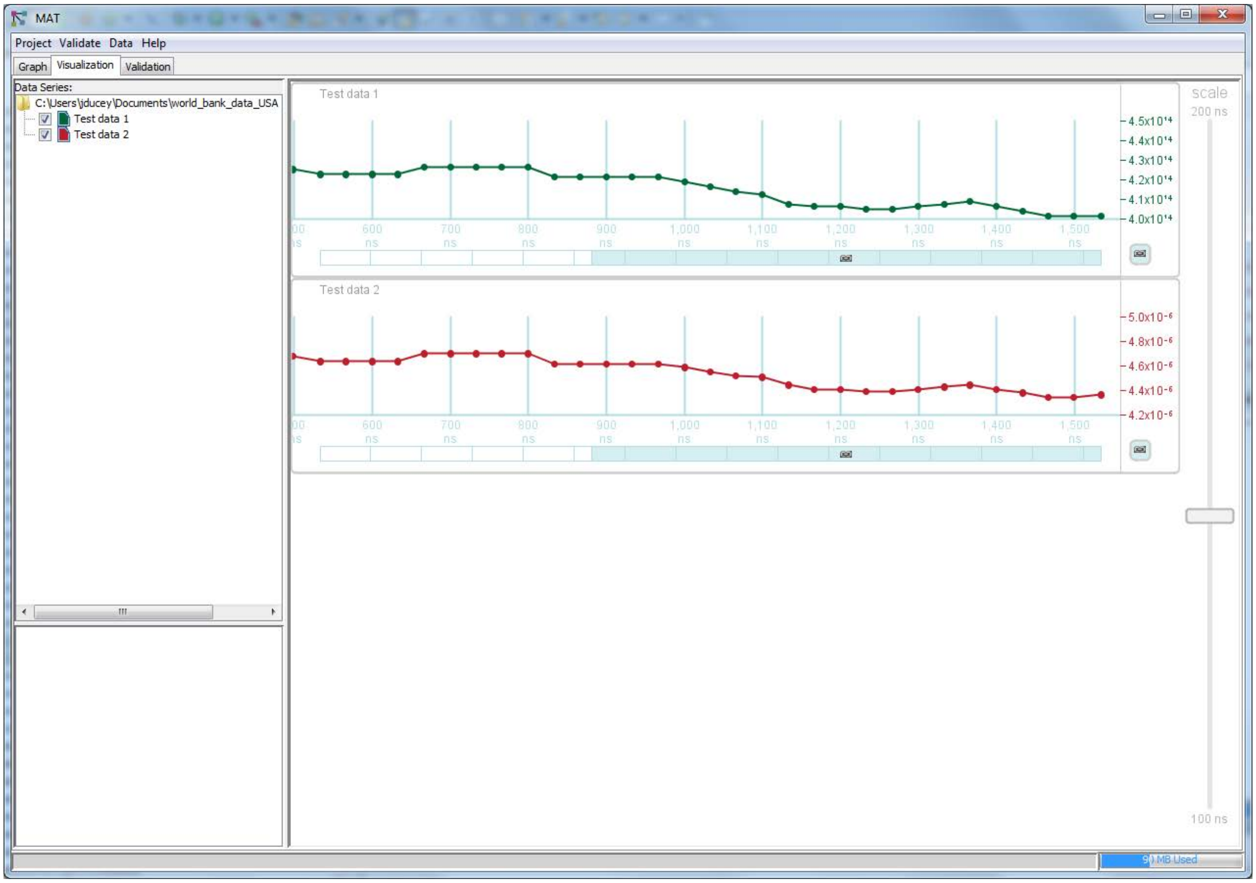
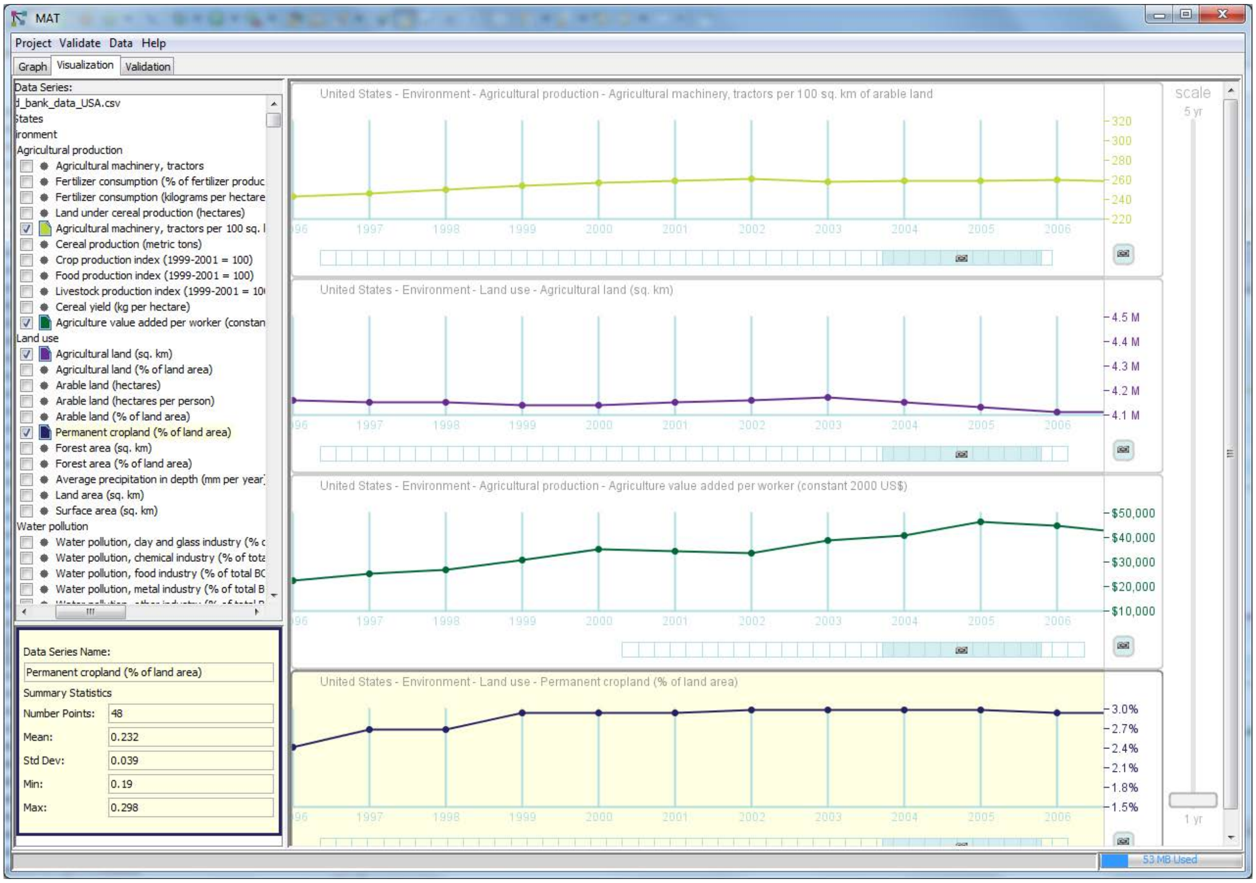


Figure 3-28: Conversion of Y-axis labels to exponential notation


MAT
Project Validate Data Help
Graph
Visualization
Validation
United States - Environment - Agricultural production - Agricultural machinery, tractors per 100 sq. km of arable land
320
300
280
260
240
220
United States - Environment - Land use - Agricultural land (sq. km)
4.5 M
4.4 M
4.3 M
4.2 M
4.1 M
United States - Environment - Agricultural production - Agriculture value added per worker (constant 2000 US$)
$50,000
$40,000
$30,000
$20,000
$10,000
United States - Environment - Land use - Permanent cropland (% of land area)
3.0%
2.7%
2.4%
2.1%
1.8%
1.5%
96
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
scale
5 yr
1 yr
53 MB Used

Data Series:
d_bank_data_USA.csv
States
ironment
Agricultural production
- ☐ Agricultural machinery, tractors
- ☐ Fertilizer consumption (% of fertilizer produc
- ☐ Fertilizer consumption (kilograms per hectare
- ☐ Land under cereal production (hectares)
- ☑ Agricultural machinery, tractors per 100 sq. l
- ☐ Cereal production (metric tons)
- ☐ Crop production index (1999-2001 = 100)
- ☐ Food production index (1999-2001 = 100)
- ☐ Livestock production index (1999-2001 = 10
- ☐ Cereal yield (kg per hectare)
- ☑ Agriculture value added per worker (constan

Land use
- ☑ Agricultural land (sq. km)
- ☐ Agricultural land (% of land area)
- ☐ Arable land (hectares)
- ☐ Arable land (hectares per person)
- ☐ Arable land (% of land area)
- ☑ Permanent cropland (% of land area)
- ☐ Forest area (sq. km)
- ☐ Forest area (% of land area)
- ☐ Average precipitation in depth (mm per year
- ☐ Land area (sq. km)
- ☐ Surface area (sq. km)

Water pollution
- ☐ Water pollution, clay and glass industry (% c
- ☐ Water pollution, chemical industry (% of tota
- ☐ Water pollution, food industry (% of total BC
- ☐ Water pollution, metal industry (% of total B

Data Series Name:
Permanent cropland (% of land area)

Summary Statistics

Number Points:	48
Mean:	0.232
Std Dev:	0.039
Min:	0.19
Max:	0.298

Figure 3-29: Examples of intelligent labeling of the Y-axis in MAT

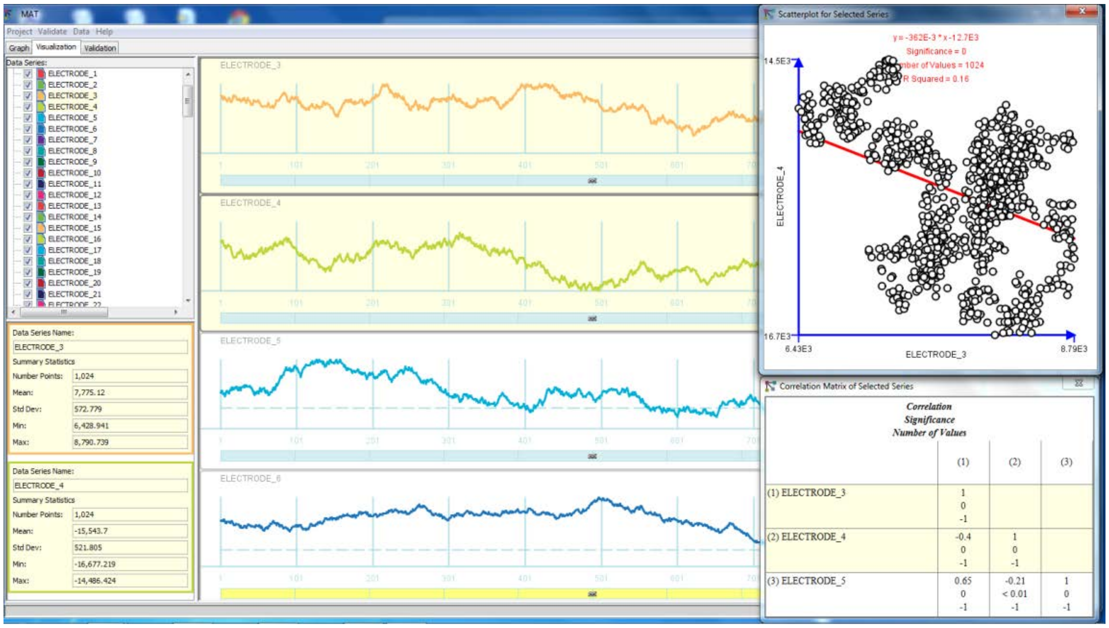
Scaling to More Data

Since MAT may be used by analysts from a variety disciplines, it is important to provide scalability in the amount of data that MAT can display. We explored various improvements to efficiently display data within MAT. These included using hardware-based graphics renderers (i.e., OpenGL), avoiding slower rendering commands in favor of faster ones, and selectively drawing subsets of the data if there is no effect on the resulting visualization. These speed improvements make it possible to visualize and interact with larger amounts of data. We will continue to consider how to most efficiently visualize data during future development to make MAT useful to the broadest audience possible.

3.4.2. New Domains

We identified an initial test case—the evaluation of Electroencephalography (EEG) data. EEG is gathered using a mesh of sensors that record electrical activity on the scalp that is indicative of neural activity. EEG data is time series data used to understand the human brain, so is a practical use case for MAT. In addition, because individual sensors can be applied improperly, throwing off the data for that sensor, it has the potential to be a useful test case for our data validation effort.

We found freely available simulated EEG data in CSV format at http://www.grid-tools.com/resources/test_data_sets.php. We also found EEG data from an experiment of rats' reactions to stimuli and from a Grand Mal seizure at http://www.vis.caltech.edu/~rodri/data.htm that was in a text format that we could convert into CSV format. We imported these data sources into MAT for exploration and evaluation. Figure 3-30, Figure 3-31, and Figure 3-32 show these data sources being analyzed in MAT.



Data Series Name: ELECTRODE_3
Summary Statistics
Number Points: 1,024
Mean: 7,775.12
Std Dev: 572.779
Min: 6,428.941
Max: 8,790.739

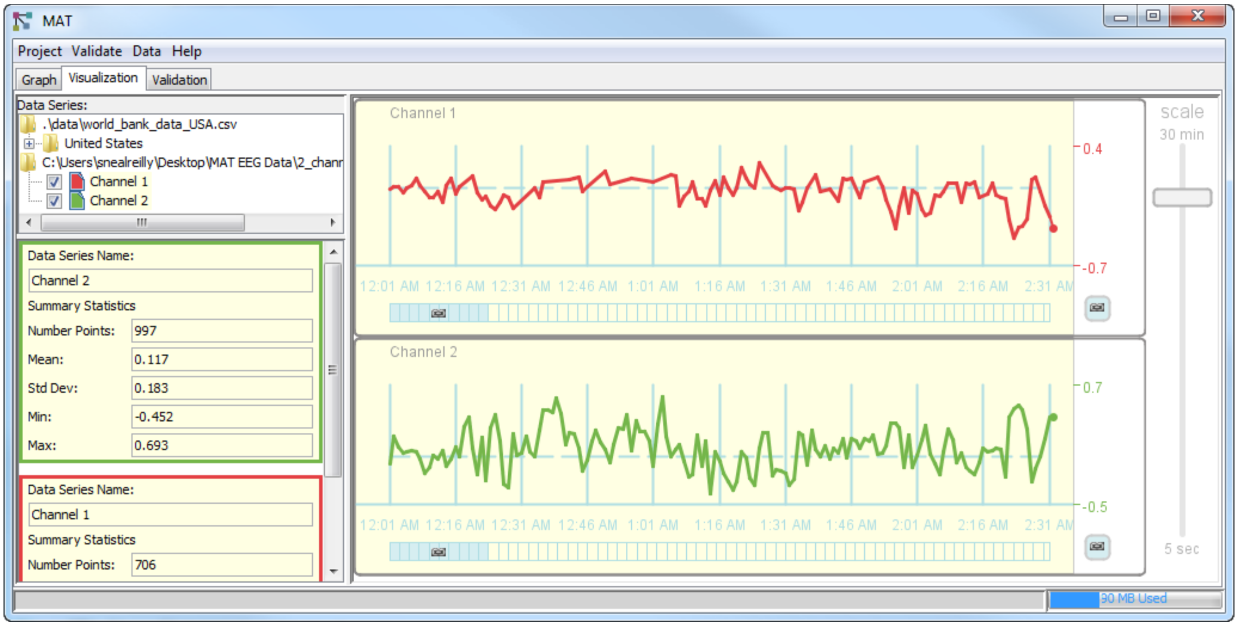
Data Series Name: ELECTRODE_4
Summary Statistics
Number Points: 1,024
Mean: -15,543.7
Std Dev: 521.805
Min: -16,677.219
Max: -14,486.424

y = -362E-3 * x -12.7E3
Significance = 0
Number of Values = 1024
R Squared = 0.16

Correlation Matrix of Selected Series

Correlation *Significance* *Number of Values*	(1)	(2)	(3)
(1) ELECTRODE_3	1 0 -1		
(2) ELECTRODE_4	-0.4 0 -1	1 0 -1	
(3) ELECTRODE_5	0.65 0 -1	-0.21 < 0.01 -1	1 0 -1

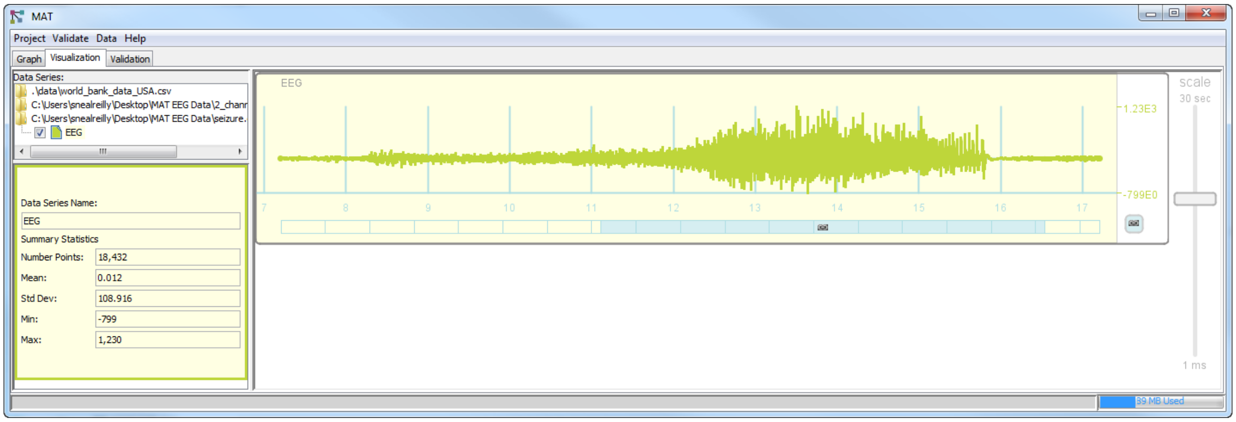
Figure 3-30: Simulated EEG data being analyzed in MAT



Data Series Name: Channel 2
Summary Statistics
Number Points: 997
Mean: 0.117
Std Dev: 0.183
Min: -0.452
Max: 0.693

Data Series Name: Channel 1
Summary Statistics
Number Points: 706

Figure 3-31: Rat EEG data being analyzed in MAT



Data Series:
.\data\world_bank_data_USA.csv
C:\Users\snealreilly\Desktop\MAT EEG Data\2_chann
C:\Users\snealreilly\Desktop\MAT EEG Data\seizure.
☑ EEG

Data Series Name: EEG

Summary Statistics

Number Points:	18,432
Mean:	0.012
Std Dev:	108.916
Min:	-799
Max:	1,230

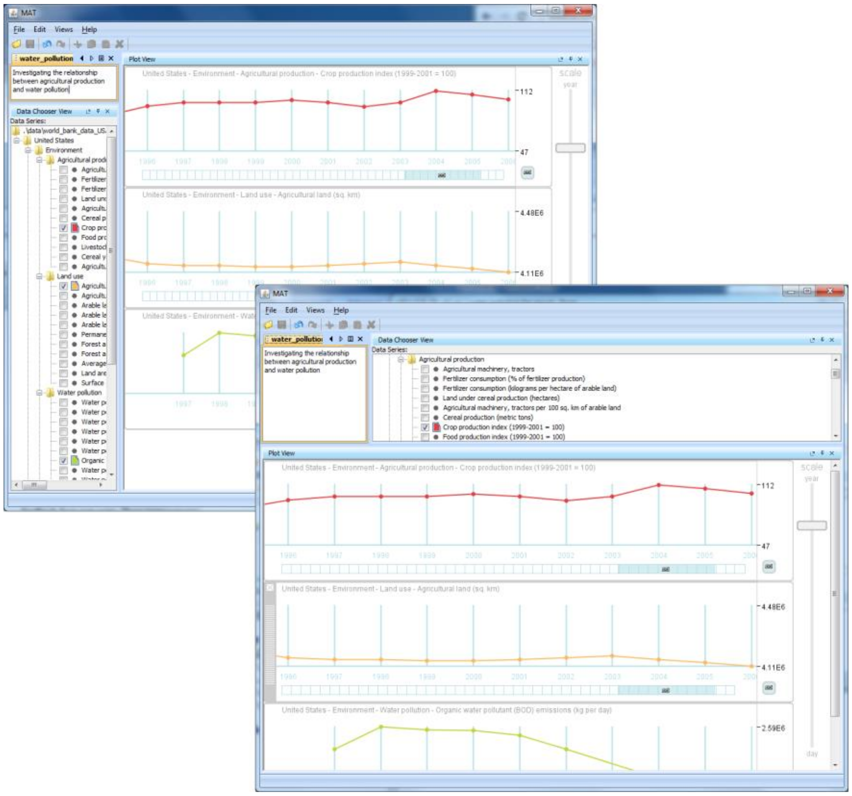
Figure 3-32. EEG data from a Grand Mal seizure

3.5. Software Improvements

3.5.1. Improvements to Usability

One of our ongoing goals is to ensure that the MAT tool is actually useful to and usable by the scientific community. So, while most of this effort is focused on novel science, we are also constantly striving to ensure that the MAT software provides a usable framework for getting this science out to the community. To this end, we are improving the user interface for MAT with a focus on more customizability and better support for analyzing data and models across multiple types of analysis. For example, we enable users to view the validation and data analysis panes at the same time and to highlight and center on a feature in the validation pane when it is clicked on in the data visualization pane.

We improved the graphical user interface by using Charles River's Metronome framework. This framework is built on top of the same Equinox libraries that the popular Eclipse Development Environment uses. In addition to increased robustness, the framework provides facilities for rearranging user interface components, providing the user with more flexibility when using MAT. For example, if the names of the data series are long, then they could be cut off in the user interface, but after rearranging the components, the names are fully visible (see Figure 3-33). The Metronome framework also provides functionality for undo and redo, so the user can easily correct mistakes.



Investigating the relationship between agricultural production and water pollution

Data Series:

- Agricultural production
  - ☐ Agricultural machinery, tractors
  - ☐ Fertilizer consumption (% of fertilizer production)
  - ☐ Fertilizer consumption (kilograms per hectare of arable land)
  - ☐ Land under cereal production (hectares)
  - ☐ Agricultural machinery, tractors per 100 sq. km of arable land
  - ☐ Cereal production (metric tons)
  - ☑ Crop production index (1999-2001 = 100)
  - ☐ Food production index (1999-2001 = 100)

United States - Environment - Agricultural production - Crop production index (1999-2001 = 100)

United States - Environment - Land use - Agricultural land (sq. km)

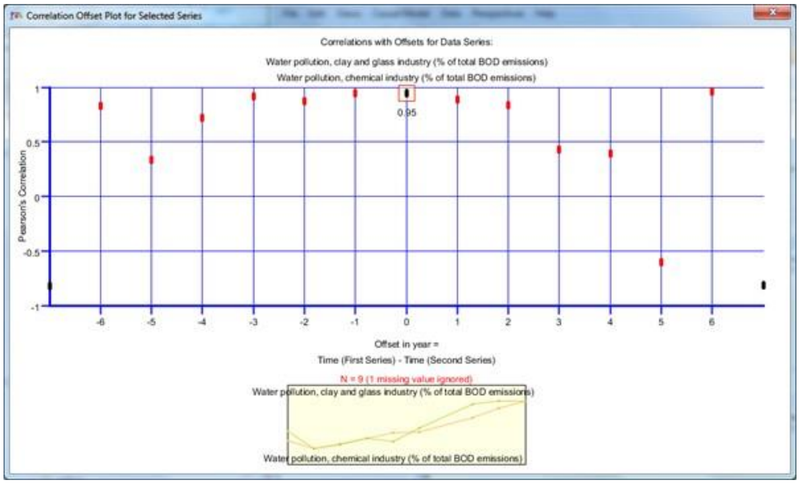
United States - Environment - Water pollution - Organic water pollutant (BOD) emissions (kg per day)

Figure 3-33: Changing pane sizes and layouts in the new Metronome-enhanced MAT

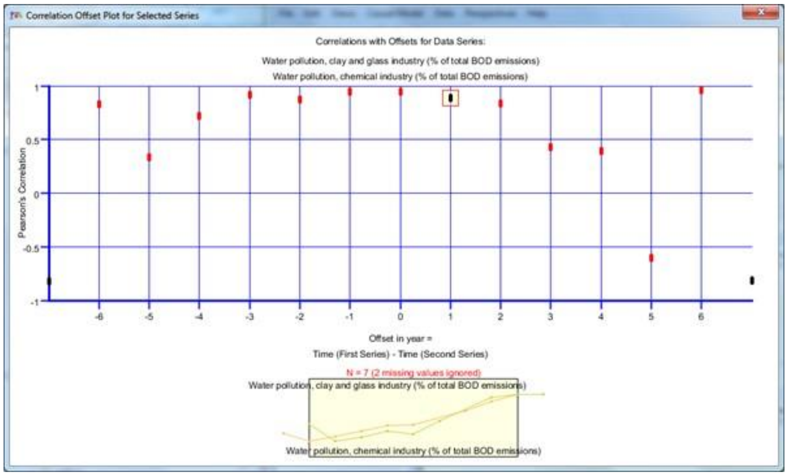
We also improved the MAT project file format so that changes in the user interface (e.g., color and layout of a data series) can be persisted after the MAT application is closed.

3.5.2. Improvements in Handling Missing Values

To provide a clearer picture to the user, it is important to provide information about what data was used in an analysis. It is common for data to have missing values that could influence the results, so improvements were made to show how many data points were used in an analysis. Furthermore, some analyses have a temporal offset as input which can change the number of values used for a calculation. For example, when performing the correlation of two time series with a temporal offset, the following screenshot lets the user know that with no temporal offset that nine points were used in the correlation calculation and there was one missing value.


Correlation Offset Plot for Selected Series
Correlations with Offsets for Data Series:
Water pollution, clay and glass industry (% of total BOD emissions)
Water pollution, chemical industry (% of total BOD emissions)
Pearson's Correlation
1
0.5
0
-0.5
-1
0.95
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
Offset in year =
Time (First Series) - Time (Second Series)
N = 9 (1 missing value ignored)
Water pollution, clay and glass industry (% of total BOD emissions)
Water pollution, chemical industry (% of total BOD emissions)

By increasing the temporal offset to one, we now have two missing values.


Correlation Offset Plot for Selected Series
Correlations with Offsets for Data Series:
Water pollution, clay and glass industry (% of total BOD emissions)
Water pollution, chemical industry (% of total BOD emissions)
Pearson's Correlation
1
0.5
0
-0.5
-1
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
Offset in year =
Time (First Series) - Time (Second Series)
N = 7 (2 missing values ignored)
Water pollution, clay and glass industry (% of total BOD emissions)
Water pollution, chemical industry (% of total BOD emissions)

3.5.3. Improvements in Handling Large Datasets

To increase rendering speed so that large datasets can be displayed in MAT, the newest stable release of the graphics library used by MAT (Processing) was introduced. We also added support for hardware acceleration and added a "smart" downsampling mechanism such that not all data points are drawn on the screen—only those that need to be given the screen resolution. When viewing datasets with hundreds of thousands of points, only a small subset need to be drawn in practice.

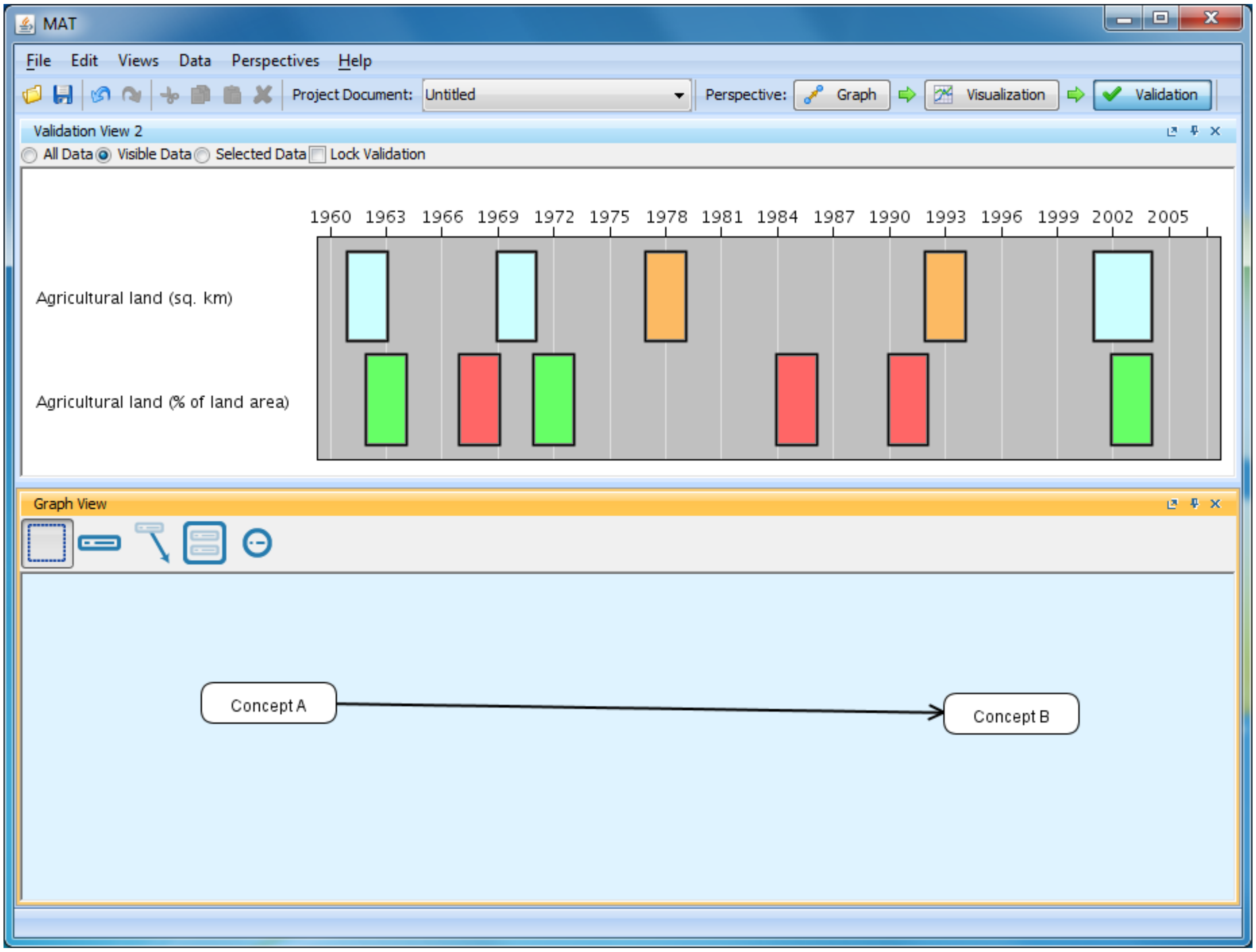
Overall, while exact performance depends on the machine's speed, these improvements allowed for better handling of datasets with millions of data points. This will help make MAT useful in a broader array of domains.

3.5.4. Data Synthesis Tool

To identify non-trivial predictive/causal relationships in data when the relationship is not between two raw data streams, but between manipulations of those streams, we developed a data synthesis tool. For instance, variable X might increase with the log of variable Y. Standard correlation analyses, even when temporal offsets are accounted for, do not find these relationships. Therefore, we added the ability to create new data series by manipulating and combining existing data series. For instance, a new data series can be created that is the average of other data series, and this new data series can be analyzed for its relationship to other events in the data. This enables MAT to express (and learn) causal models such as, "whenever the sum of the percentage of people unemployed and people who are unhappy with their job crosses a threshold…."

3.5.5. Visualizing Validation Results

The validation visualization shows whether the causal model was supported by the empirical data. As shown in Figure 3-34, the green events are effects that are supported by a cause.



Project Document: Untitled

○ All Data ● Visible Data ○ Selected Data □ Lock Validation

Figure 3-34: Events supported by a cause are displayed in green

In simple examples, it is easy to understand which cause supports each effect, but in more complex causal models, this may not be so clear. Therefore, we modified the visualization to include arrows showing the valid causal relationships. For example, in Figure 3-35, the user selected the causal relationship in the causal model, which displays the valid examples of that causal relationship as arrows pointing from cause to effect. This enables the user to quickly and easily see the relationship.

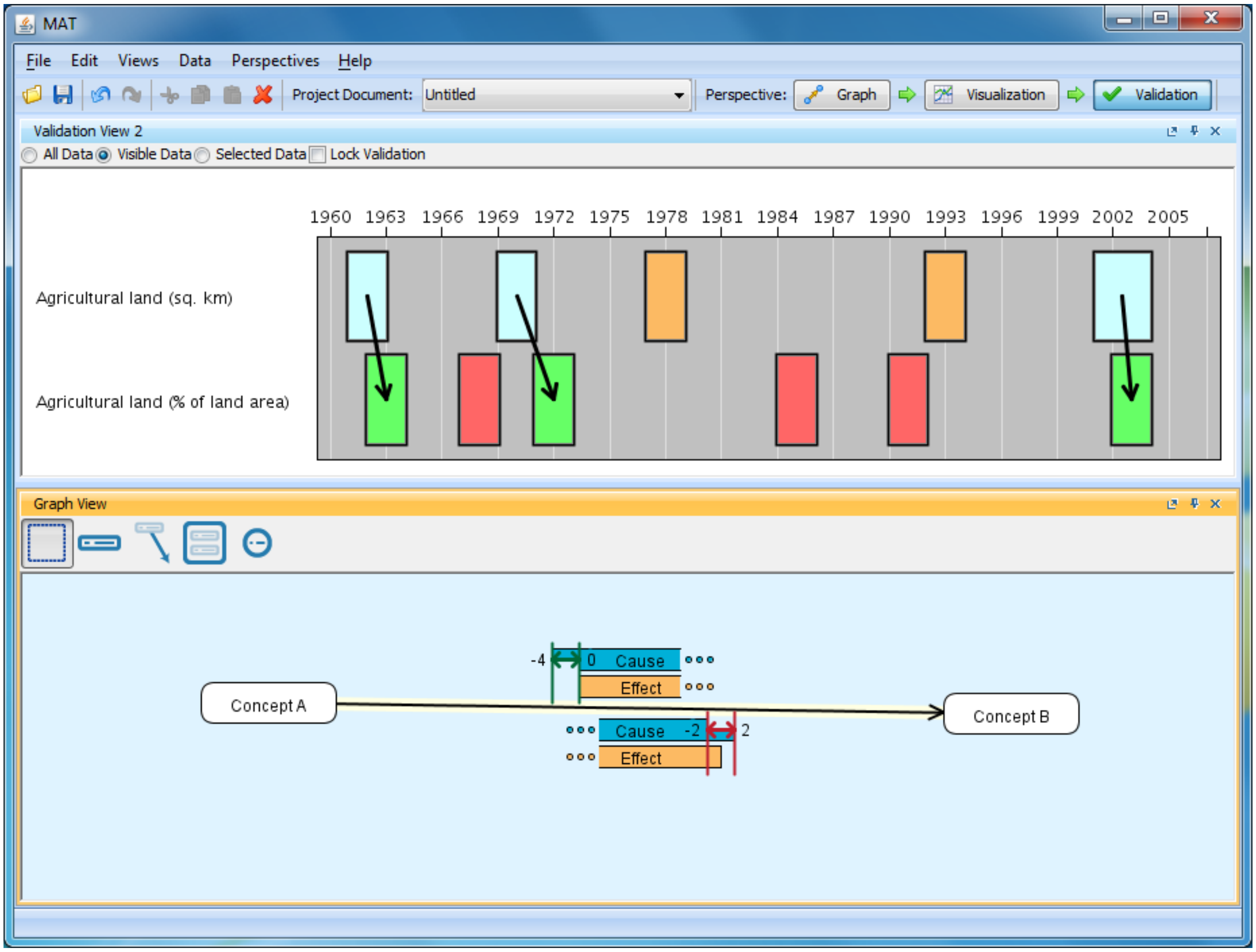
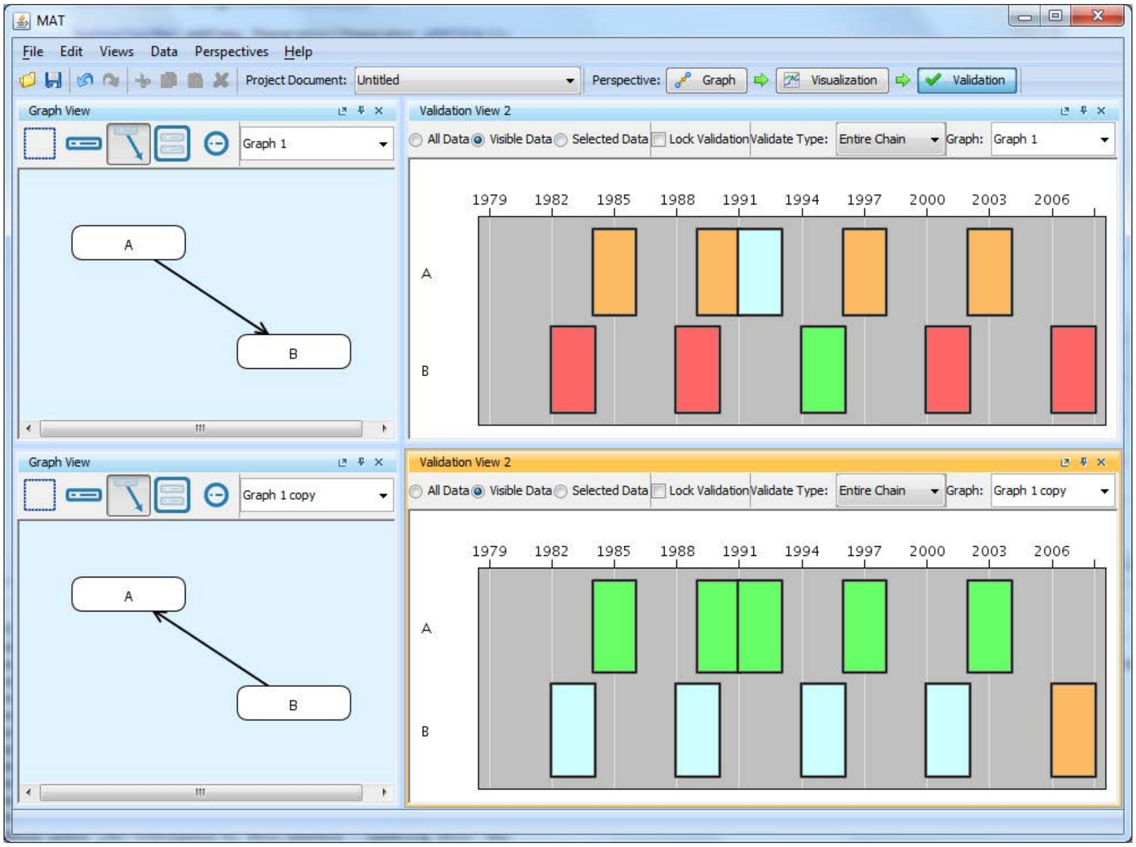

MAT
File Edit Views Data Perspectives Help
Project Document: Untitled
Perspective: Graph Visualization Validation
Validation View 2
○ All Data ● Visible Data ○ Selected Data □ Lock Validation
1960 1963 1966 1969 1972 1975 1978 1981 1984 1987 1990 1993 1996 1999 2002 2005
Agricultural land (sq. km)
Agricultural land (% of land area)
Graph View
Concept A
Concept B
-4
0
Cause
Effect
Cause
-2
2
Effect

Figure 3-35: Causes and effects are connected with arrows

3.5.6. Comparison of Competing Models

The causal relationship between two concepts is not always clear. Therefore, MAT enables the user to create multiple causal models for comparison. The screenshot in Figure 3-36 shows a simple example where the user is unsure whether Concept A causes B or vice versa. Both causal models can be created and validated using the same empirical data. In this example, it is clear that the causal model where Concept B causes A is the better of the two models.



Project Document: Untitled

Graph 1

Validate Type: Entire Chain

Graph: Graph 1

Graph 1 copy

Validate Type: Entire Chain

Graph: Graph 1 copy

Figure 3-36: Model comparison user interface

The causal relationship can be explored by selecting the nodes in the graph to show the valid causal relationships. Figure 3-37 shows the MAT screenshot that demonstrates this relationship. An important challenge when including multiple causal models is to ensure that it is clear to the user when a node in a different model represents the same concept. Therefore, when a concept is selected in one causal model, it also appears as selected in all other causal models that use this concept. The management of concepts in an intuitive way, where the user has both concepts common across models and unique to a single model, is under development.

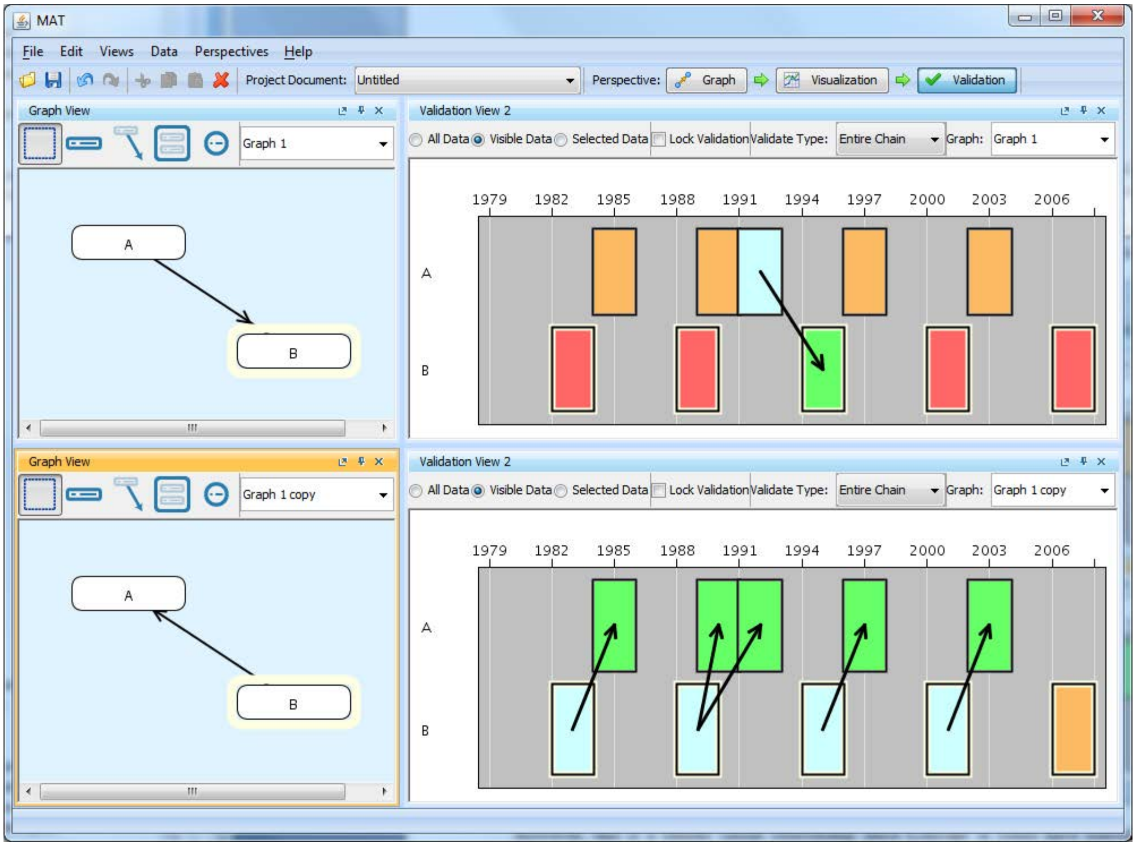
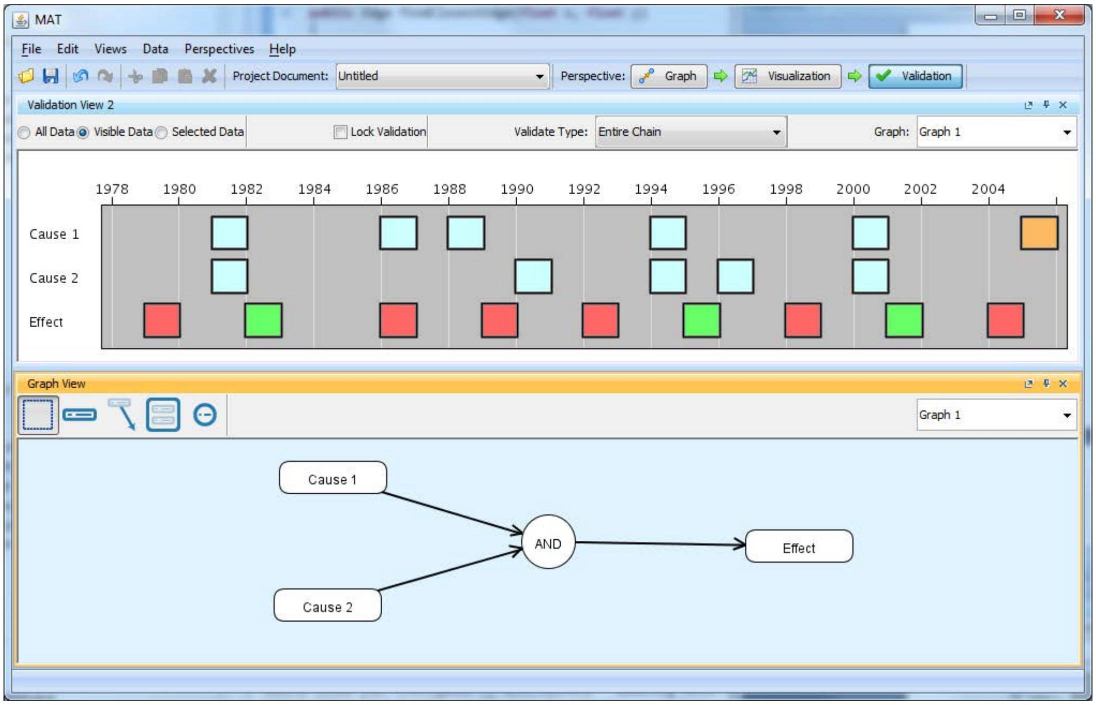


Figure 3-37: Node B selected to show causal relationships

3.5.7. Exploration of Causal Models Using Undo and Redo

To encourage the exploration of causal models, actions performed on the models can be undone. Figure 3-38 shows the MAT screenshot with the results after validating the model with data. In the bottom row of the timeline chart, the green features are supported by evidence, while the red ones are not.



Project Document: Untitled

Validate Type: Entire Chain

Graph: Graph 1

Graph 1

Figure 3-38: Validated model showing evidence for features

If the user wanted to see how the model changed using an OR logic node instead of an AND logic node, they can change that node. Figure 3-39 shows the automatic update of the validation.

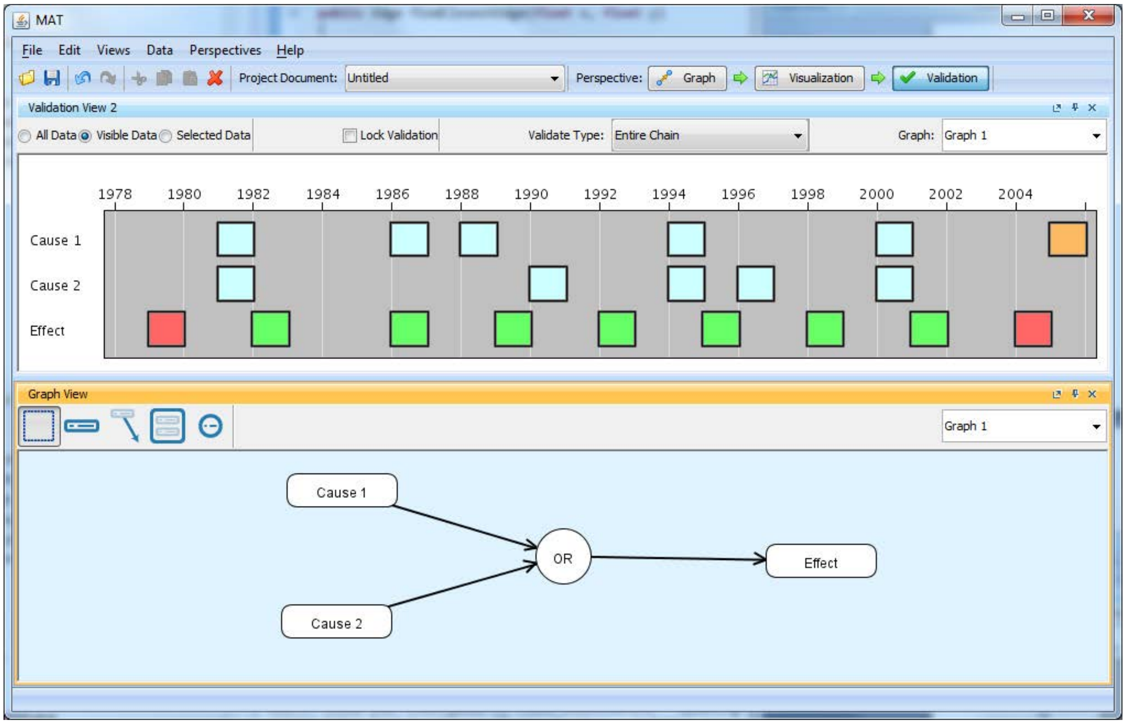


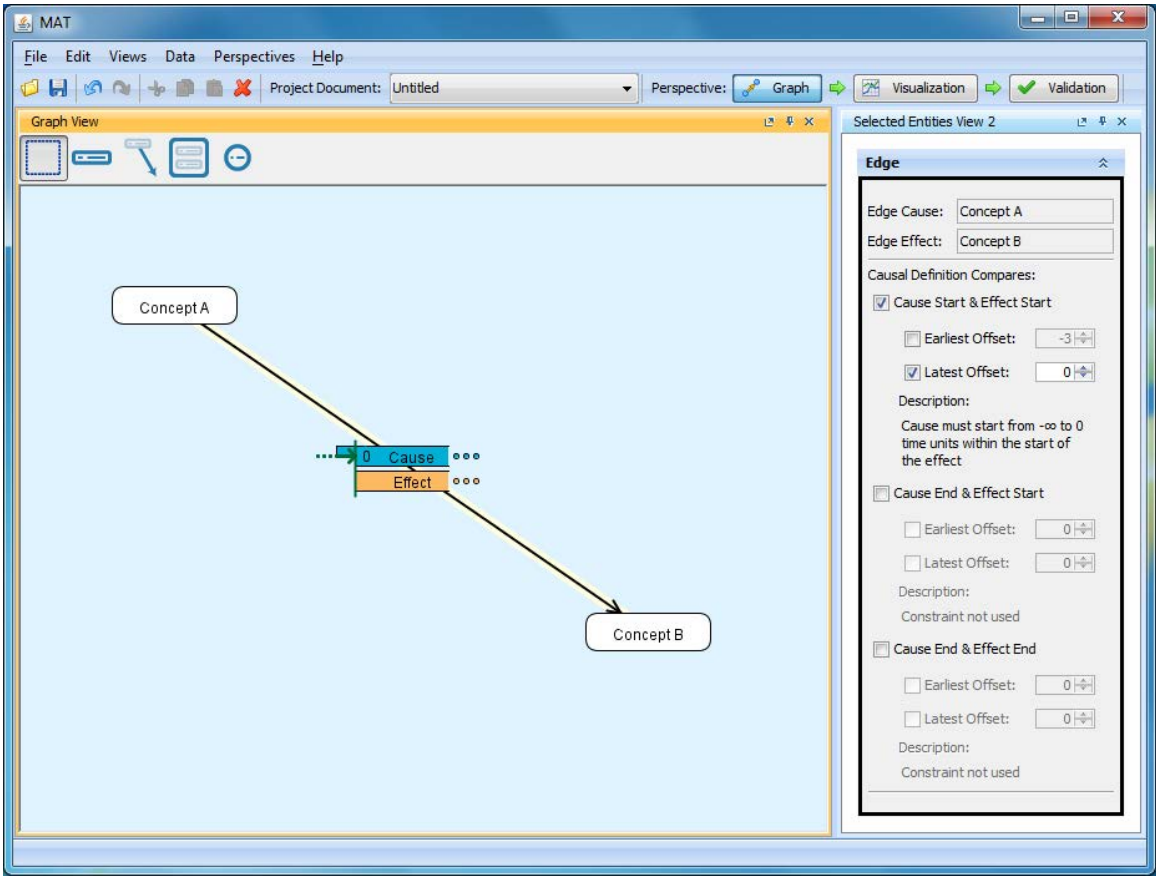
Figure 3-39: Node changed from AND to OR and results on evidence

The undo and redo operators in the toolbar enables the user to switch back and forth between the previous and current version of the causal model. The user can continue to make adjustments to the model to see how it changes the data validation and then use the undo operator to return to the original version of the model. This provides another way for the user to explore their causal models and make model adjustments to see the influence on validation results.

3.5.8. Graphical Tool for Specifying Constraints on Causal Relationships

One of the challenges that occur when describing causal models is how to make it possible (and, ideally, easy) to specify the temporal constraints on a causal link. For instance, if the hypothesis is that A (e.g., unemployment) leads to B (e.g., crime), and the data indicates that there is a jump in crime after a spike in unemployment, the temporal distance between A and B effects whether we want to associate the two events. If B happens 6 months after A, it could be counted as evidence. It B happens 6 decades after A, we probably want to treat them as unrelated events. Specifying these temporal constraints is difficult—the relationships are not simply "starts-after," but can relate to other factors, such as when A or B ended. Our previous mechanism for specifying these constraints was found to be unintuitive and users requested additional functionality, so we updated and expanded this functionality.

To make the creation of causal models as easy as possible, a new visualization of the causal relationships was developed. Now, the user can see how the cause and effect are compared when validating a causal relationship. For example, in the causal model in Figure 3-40, Concept A causes Concept B using the simplest of causal relationships where Concept A must start sometime before Concept B.



Project Document: Untitled

Edge

Edge Cause: Concept A

Edge Effect: Concept B

Causal Definition Compares:

☑ Cause Start & Effect Start

□ Earliest Offset: -3

☑ Latest Offset: 0

Description:

Cause must start from -∞ to 0 time units within the start of the effect

□ Cause End & Effect Start

□ Earliest Offset: 0

□ Latest Offset: 0

Description:

Constraint not used

□ Cause End & Effect End

□ Earliest Offset: 0

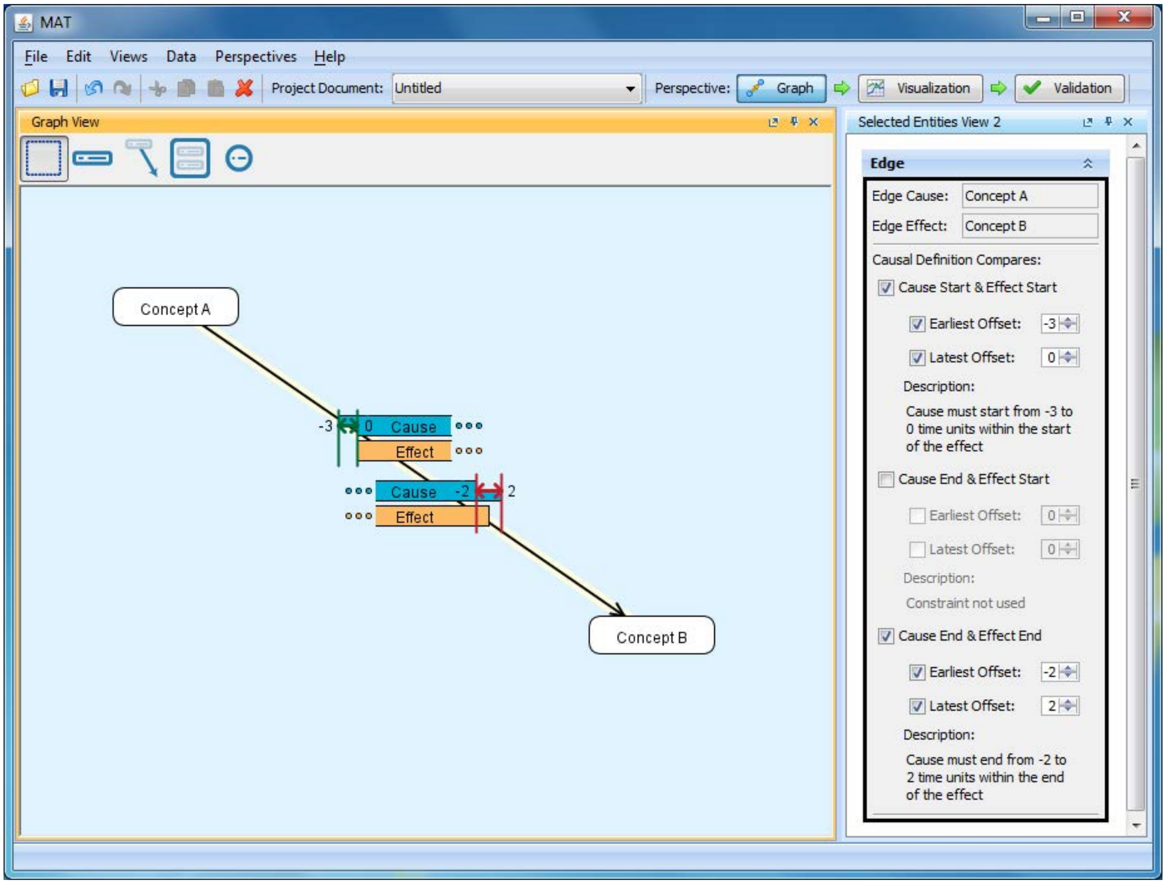
□ Latest Offset: 0

Description:

Constraint not used

Figure 3-40: Causal model showing Concept A starting before Concept B

In this causal relationship, Concept A can start and end many years (or centuries) before Concept B. Therefore, a more useful causal relationship includes a window that restricts when the cause starts relative to the start of the effect. Also, a second type of constraint can be included that restricts when the cause ends relative to the end of the effect. Figure 3-41 shows this new causal relationship.



Project Document: Untitled

Edge

Edge Cause: Concept A

Edge Effect: Concept B

Causal Definition Compares:

☑ Cause Start & Effect Start

☑ Earliest Offset: -3

☑ Latest Offset: 0

Description:

Cause must start from -3 to 0 time units within the start of the effect

□ Cause End & Effect Start

□ Earliest Offset: 0

□ Latest Offset: 0

Description:

Constraint not used

☑ Cause End & Effect End

☑ Earliest Offset: -2

☑ Latest Offset: 2

Description:

Cause must end from -2 to 2 time units within the end of the effect

Figure 3-41: Model specifying start and end windows for cause and effect

The panel on the right enables the user to add, remove, and modify three types of causal constraints that compare:

- Cause and effect start
- Cause and effect end
- Cause end and effect start

The panel on the right provides a verbal description of each of the causal constraints to further improve the user's understanding of the causal relationship.

3.5.9. Handling Ambiguous Causal Models

When an effect in the causal model has two possible causes, the user's intentions can be ambiguous. For example, Figure 3-42 shows a causal model with two possible causes, but it is unclear if both causes must be present for the effect to occur or if one cause is sufficient.

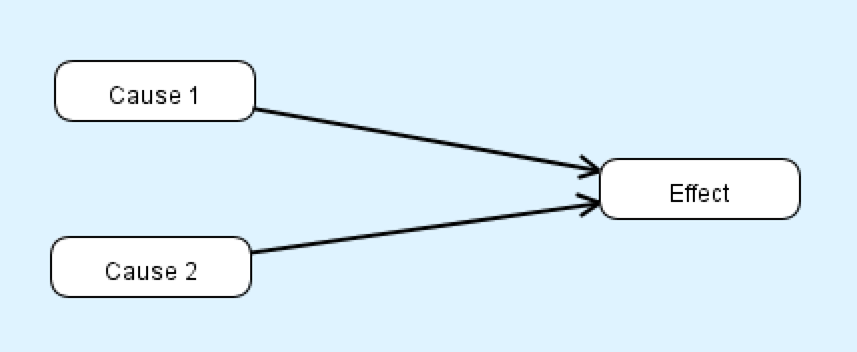
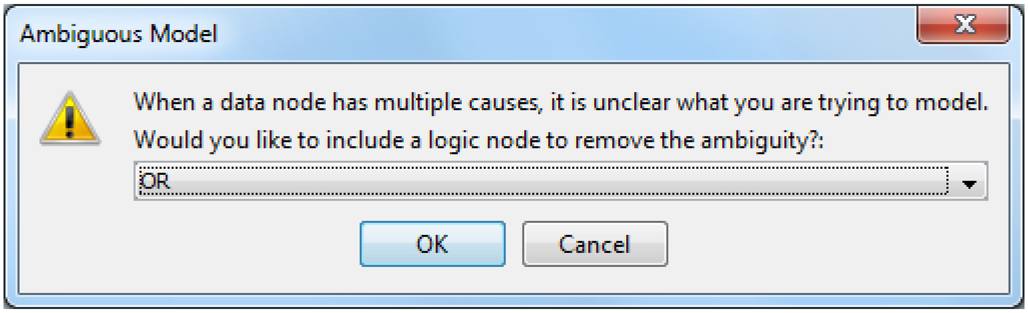


Figure 3-42: Ambiguous cause for an effect

Therefore, a warning message appears when the user includes a second cause to an effect. Figure 3-43 shows the warning message that notifies the user of the ambiguity.



When a data node has multiple causes, it is unclear what you are trying to model.
Would you like to include a logic node to remove the ambiguity?:

OR

Figure 3-43: Warning message

The user can then specify the intended relationship and the necessary logic will be put in place. Figure 3-44 shows the specified relationship.

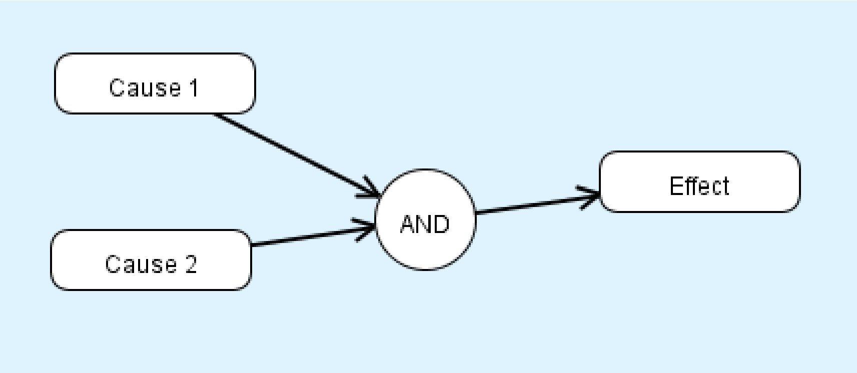


Figure 3-44: Disambiguated causal relationship

If the model was changed in a way that was not the user’s intention, then the user can undo the operation and go back to the previous state of the causal model.

3.6. Evaluation and Transition

We focused throughout the program on making MAT available to the government and academic research communities and to look for opportunities to use MAT on a variety of ongoing research efforts. We provided copies of MAT to the following institutions based on their requests for the software: the University of Michigan, Arizona State University, Kansas State University, University of California at Los Angeles, the Naval Medical Research Unit at Wright Patterson Air Force Base, Concordia University (Montreal), the University of Wisconsin, the University of Maryland, and the Air Force Research Laboratory's Human Effectiveness Directorate, the Intelligence Advanced Research Projects Agency (IARPA), the Joint Advanced Warfighting Division (JAWD), Sandia National Labs, Los Alamos National Laboratory, and Lockheed Martin's ICEWS team.

We also published the following papers over the course of the project:

- Amy Sliva, Scott Neal Reilly, "A Big Data Methodology for Bridging Quantitative and Qualitative Political Science Research," American Political Science Association (APSA) 2014, Washington, DC (2014)
- Amy Sliva, Scott Neal Reilly, Randy Casstevens, John Chamberlain, "Tools for Validating Causal and Predictive Claims in Social Science Models," 6th International Conference on Applied Human Factors and Ergonomics (AHFE) 2015, Las Vegas, NV (2015)
- Amy Sliva, Scott Neal Reilly, Randy Casstevens, John Chamberlain, "Validating Causal and Predictive Claims in Sociocultural Models" in Denise Nicholson, CDR Joseph Cohn, LT David Combs, and Sae Schatz (eds.), "Modeling sociocultural influences on decision making" (Forthcoming)

As a final note, MAT received a nomination from the American Political Science Association (APSA) for Best Statistical Software of 2015. The awards are to be made in September 2015.

4. Recommendations for Future Research

Understanding causal relationships in the world is one of the fundamental quests of mankind. It is only by understanding how the world (the physical world, the social world, etc.) works *causally* that we can effect change—whether that change is curing diseases, reducing violence and poverty, or engaging in successful military missions. Many causal relationships are known and well understood, but others are much more subtle, complex (e.g., in the case of socio-political relationships), or, especially when dealing with adversaries, obscure. In hard-science domains, we are able to use controlled experiments to identify and tease apart causal relationships in the world. Soft sciences (e.g., politics, social science, economics), however, do not typically have this luxury. Similarly, military planning domains, including for both kinetic and non-kinetic operations, also fail to provide opportunities for controlled experiments. That does not mean that understanding causal processes is impossible in these domains; we just need to develop more sophisticated tools to help extract and validate causal theories using observational data where we cannot rely on the existence of experimental data.

We have developed the Model Analyst's Toolkit (MAT) to help build, refine, and validate causal and predictive models of a wide range of phenomena. For instance, MAT has been used on a range of models from socio-political models to physiological models to business models. One of the fundamental gaps in MAT's capabilities, however, is that MAT analyzes time-series data to reason about causal relationships. This is obviously useful when such data is available, but we have found that more often than not it is not available. For instance, social scientists gathering survey data will often gather data from many people, and sometimes from many places, but often at only a single point in time (or at a small number of points in time). If we want to understand, say, whether joining a particular group or being in prison causes radicalization or whether radicalization leads to joining particular groups or spending time in prison, simple correlational models will not help. If we had temporal data of an individual's level of radicalization over time and could relate that to when they joined a particular radical group or spent time in prison, we could start to tease apart cause from effect. But if all we know is the state of affairs at a given point in time, understanding which are the causes and which are the effect becomes much more challenging—but since most data is of this form, it is also that much more useful.

Even in domains like medicine, the long-term longitudinal studies necessary for sound time-series based analyses are rare and, by definition, require significant amounts of time before they provide results. Instead, it is common for observational studies that look at correlations between, say, heart health and drinking red wine to drive academic articles, press reports, and the behavior of many people, even though the causal links (and, often, explicit or implicit claims of such causal links) in such observational studies are not rigorously evaluated. In addition, in domains like medicine (and other experimental sciences), using observational data to find likely causal relationships can be a step towards designing formal experiments that focus on gaining a deeper understanding of phenomena that are first identified in the observational data.

If we could extend the functionality of MAT to support the causal analysis of data that is not in time-series form, and combine this with the existing temporal analyses, we believe we can dramatically increase the usefulness of MAT. We also believe that the scientific foundations that will enable this sort of extension are recently becoming available. For instance, recent developments coming out of the field of Uncertainty in AI show significant promise for non-temporal causal analysis. In addition, insights from ensemble machine learning and mixed initiative systems can be applied in novel ways to the problem of reasoning about causality from observational data, combining the strengths of a variety of automated analysis techniques and human expertise to achieve improved results. Not only will this enhanced software workbench open the door for researchers to assess causal relationships in their non-time-series data, but it will also enable them to combine evidence from a range of temporal and non-temporal datasets into a single, rich representation of causal processes.

Some of the questions we believe should be addressed by future research include:

- Can we use automated analysis techniques to support the construction of causal models to explain and forecast events? Can we do this when only non-time-series observational data is available? Can we combine results when time-series and non-time-series data is available?

- Can we use automated analysis techniques to support the validation of causal models from observational data (in time-series and non-time-series forms)? Can we use automated analysis techniques to support the extraction of causal links from observational data?
- Can we combine different causal modeling techniques to improve these analyses? Can we learn to do so better over time?
- Can we create a mixed-initiative, user-focused software system to support the effective use of these analysis capabilities?

5. Summary of Costs

Our total budget was $928,224. While final costs remain to be completed, the breakdown of costs at this point in the program (which given the late date is necessarily close to the final numbers), is roughly 1% for travel (and associated burdening) and 99% for labor (including associated burdening and our contractor developer). There were no material costs on this project.

Overall, we are happy to report that we completed the project on time and within budget.

6. References

Braithwaite, Alex, Niheer Dasandi, and David Hudson (2014). "Does poverty cause conflict? Isolating the causal origins of the conflict trap." Conflict Management and Peace Science.

Collier, Paul; Elliott, V. L.; Hegre, Håvard; Hoeffler, Anke; Reynal-Querol, Marta; Sambanis, Nicholas. (2003). Breaking the Conflict Trap: Civil War and Development Policy. Washington, DC: World Bank and Oxford University Press.

Collier, P. and Hoeffler, A. (2004). "Greed and Grievances in Civil Wars." Oxford Economic Papers 56.

Dereszynski, E. W.& Dietterich, T. G. (2011). Spatiotemporal models for data-anomaly detection in dynamic environmental monitoring campaigns. ACM Transactions on Sensor Networks (TOSN), 8.

Dereszynski, E.& Dietterich, T. (2007). Probabilistic models for anomaly detection in remote sensor data streams.

Djankov, S. and Reynal-Querol, M. (2008). “Poverty and Civil War: Revisiting the evidence”, CEPR No. 6980.

Neal Reilly, S. (2010). Validation Coverage Toolkit for HSCB Models. Charles River Analytics Inc.

Neal Reilly, W. S., Pfeffer, A., and Barnett, J. (2010). A Metamodel Description Language for HSCB Modeling. In Advances in Cross-Cultural Decision Making.

Neal Reilly, W. S., Pfeffer, A., Barnett, J., Chamberlain, J., and Casstevens, R. (2011). A Computational Toolset for Socio-Cultural Data Exploration, Model Refinement, and Model Validation. In Human Social Culture Behavior (HSCB) Focus.

The World Bank. World Development Indicators. 2013.

Themnér, Lotta & Peter Wallensteen (2014) Armed Conflict, 1946-2013. Journal of Peace Research 51(4).